# Faster Subset Selection for Matrices and Applications

**Haim Avron**
Business Analytics & Mathematical Sciences
IBM T.J. Watson Research Center
haimav@us.ibm.com

**Christos Boutsidis**
Business Analytics & Mathematical Sciences
IBM T.J. Watson Research Center
cboutsi@us.ibm.com

September 26, 2012

**Abstract**

We study the following problem of *subset selection* for matrices: given a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ ($m > n$) and a sampling parameter $k$ ($n \leq k \leq m$), select a subset of $k$ columns from $\mathbf{X}$ such that the pseudo-inverse of the sampled matrix has as smallest norm as possible. In this work, we focus on the Frobenius and the spectral matrix norms. We describe several novel (deterministic and randomized) approximation algorithms for this problem with approximation bounds that are optimal up to constant factors. Additionally, we show that the combinatorial problem of finding a low-stretch spanning tree in an undirected graph corresponds to subset selection, and discuss various implications of this reduction.

## 1 Introduction

Given a full rank short-and-fat matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ with $m > n$ (typically $m \gg n$) it is often of interest to *compress* $\mathbf{X}$ via selecting a subset of its columns. The goal of such a sampling procedure is to select the columns in a way that the sampled matrix behaves spectrally similarly to the original matrix, i.e. the singular values of the two matrices are comparable. Since deleting columns from $\mathbf{X}$ decreases the singular values monotonically (this is immediate from the interlacing property of the singular values; see Theorem 8.1.7 on page 396 in [29]), the challenge is to select the columns that maximize the spectrum in the sampled matrix. One can naturally formulate this objective into the following combinatorial optimization problem (let $[m] = \{i \in \mathbb{N} : i \leq m\}$, i.e. the set of natural numbers $1, 2, ...m$).

**Problem 1.1** (Subset Selection for Matrices)**.** *Fix $\mathbf{X} \in \mathbb{R}^{n\times m}$ with $m > n$ and a sampling parameter $k$ with $n \leq k \leq m$. Let $\mathcal{S} \subseteq [m]$ denote a set of cardinality at most $k$ and $\mathbf{X}_\mathcal{S} \in \mathbb{R}^{n\times|\mathcal{S}|}$ contain the columns of $\mathbf{X}$ indicated in $\mathcal{S}$. Among all such possible choices of $\mathcal{S}$, find an $\mathcal{S}_{opt}$ such that,*

$$\mathcal{S}_{opt} \in \arg\min_{\mathcal{S}} \|\mathbf{X}_\mathcal{S}^\dagger\|_\xi.$$

*Note that there might be more than one possibility for $\mathcal{S}_{opt}$ (the minimizer might not be unique). In the above, $\xi = 2, \mathrm{F}$ denotes the spectral or the Frobenius matrix norm, respectively, and $\mathbf{X}_\mathcal{S}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{X}_\mathcal{S}$.*

Technically, the above definition corresponds to two different combinatorial optimization problems, one for $\xi = 2$ and the other for $\xi = \mathrm{F}$.

Problem 1.1 occurs in numerous applications in combinatorial problems involving sampling: column-based low-rank matrix approximation [10, 12]; feature selection in $k$-means clustering [11, 13]; optimal experiment design [20, 21]; multipoint boundary value problems [20, 21]; sparse solutions to least-squares

regression [16, 8]; sensor selection in a wireless network [39]; rank-deficient linear least squares [27], and rank-deficient non-linear least squares [38], to name just a few. We discuss some of these applications in Section 6.

However, our initial motivation for investigating Problem 1.1 was our observation that the combinatorial problem of finding a low-stretch spanning tree in an undirected graph [2] corresponds to the Frobenius norm version of Problem 1.1. This connection is new and might be of independent interest.

We study three aspects of Problem 1.1: algorithms, lower bounds, and applications. We now summarize our contributions in each of these aspects.

**Algorithms.** In Section 3 we describe five different approximation algorithms for Problem 1.1. We suggest five different algorithms because no single algorithm has the lowest operation count; the choice of the most efficient algorithm depends on the actual values of $m$, $n$ and $k$. Our algorithms are considerably faster than previously known algorithms, and they achieve the same or tighter approximation bounds. Table 1 summarizes the algorithms we propose, as well as previously known algorithms for Problem 1.1.

Our first two algorithms, Algorithm 1 and Algorithm 2, which we describe in Theorem 3.1 and Corollary 3.3 (both in Section 3), respectively, are especially fast when $k$ is close to $m$, since they form $\mathcal{S}$ by greedily removing columns. Both algorithms are deterministic. Algorithm 1 in Theorem 3.1 is designed for the Frobenius norm case ($\xi = \mathrm{F}$). It requires $O\left(mn^2 + mn(m-k)\right)$ operations, and finds a subset $\mathcal{S}$ of cardinality $k$ such that

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \leq \frac{m-n+1}{k-n+1} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 .$$

Notice, for example, that if $k = m - \alpha$, for some small integer $0 < \alpha \leq 0.9(m-n+1)$, then the approximation bound is $1 + 10\alpha(m-n+1)^{-1}$.

Algorithm 2 in Corollary 3.3 is designed for the spectral norm case ($\xi = 2$). It's operation count is $O\left(mn^2 + mn(m-k)\right)$ as well. It finds a subset $\mathcal{S}$ of cardinality $k$ such that

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2 \leq \left(1 + \frac{n\,(m-k)}{k-n+1}\right) \cdot \|\mathbf{X}^{\dagger}\|_2^2 .$$

Similarly, if, for example, $k = n + 1 + \beta$, for some integer $\beta$ close to $m$ with $0 < \beta < m - n + 1$, then the approximation bound is $1 + n + n(m-n-1)\beta^{-1}$.

The idea of greedily removing columns was previously used by de Hoog and R. Mattheijb in [20]. However, our algorithms are at least a factor of $n$ faster, and in some cases a factor of $n^2$ faster. Furthermore, our algorithms operate on a wider range of matrices: the algorithms in [20] require that all possible column subsets in $\mathbf{X}$ of size $k$ or larger are non-singular, while our algorithms have no such restriction (see the paragraph Greedy Algorithms in Section 1.1 for a detailed discussion of the results in [20]).

Our third algorithm, which we describe as Algorithm 3 in Theorem 3.7 (Section 3), is designed for cases that $k$ is small, e.g. $k = O(n)$, a case which is common in applications (see Section 6). The algorithm's operation count is $O\left(mn^2 + kn^2m\right)$, and it constructs a subset $\mathcal{S}$ with cardinality at most $k > n$, such that, for both $\xi = 2, \mathrm{F}$:

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^2 \leq \left(1 + \sqrt{\frac{m}{k}}\right)^2 \left(1 - \sqrt{\frac{n}{k}}\right)^{-2} \|\mathbf{X}^{\dagger}\|_{\xi}^2 .$$

This algorithm is inspired by recent results on approximate decompositions of the identity [4, 10]. Notice that, for example, if $k = \Theta(n)$, the approximation bound is $1 + O(m/k)$.

Our fourth algorithm, which we describe as Algorithm 4 in Theorem 3.7 (Section 3), is designed for cases where both $m$ and $k$ are large (specifically, $k = \Omega(n \log n)$). It is especially fast since it is based on randomly sampling columns of the matrix. However, we do not use uniform sampling, so our bounds are independent of numerical properties (like coherence) of the matrix. The operation count of the algorithm is

| | **Sampling** | **Bound on** $\frac{\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2}{\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2}$ | **Bound on** $\frac{\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2}{\|\mathbf{X}^{\dagger}\|_2^2}$ | **Operation count** |
|---|---|---|---|---|
| *Old Algorithms* | | | | |
| Theorem 2 in [20] | $k \geq n$ | $\frac{m-n+1}{k-n+1}$ | $\frac{m-n+1}{k-n+1} \cdot n$ | $O(mn^3(m-k))$ |
| Corollary 2 in [20] | $k \geq n$ | $\frac{m-n+1}{k-n+1} \cdot \frac{n\|\mathbf{X}^{\dagger}\|_2^2}{\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2}$ | $1 + \frac{n(m-k)}{k-n+1}$ | $O(mn^3(m-k))$ |
| Theorem 1 in [21] | $k \geq n$ | $\frac{m-n+1}{k-n+1} \cdot \frac{n\|\mathbf{X}^{\dagger}\|_2^2}{\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2}$ | $1 + \frac{n(m-k)}{k-n+1}$ | $O(mn^3(m-k))$ |
| Lemma 16 in [9] | $k = n$ | No bound | $1 + f^2 n(m-n)$ | $O(mn^2 \log_f m)$ |
| Lemma 1 in [28] $\delta = 1/2$ | $k > 8 \cdot \tau \cdot n \cdot \log(2n)$ | No bound | $\frac{2m}{k}$ | $O(mn^2 + k)$ |
| Section A in [39] | $k \geq n$ | No bound | No bound | $O(m^3)$ |
| *New Algorithms* | | | | |
| Theorem 3.1 | $k \geq n$ | $\frac{m-n+1}{k-n+1}$ | $\frac{m-n+1}{k-n+1} \cdot n$ | $O(mn^2 + mn(m-k))$ |
| Corollary 3.3 | $k \geq n$ | $\frac{m-n+1}{k-n+1} \cdot \frac{n\|\mathbf{X}^{\dagger}\|_2^2}{\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2}$ | $1 + \frac{n(m-k)}{k-n+1}$ | $O(mn^2 + mn(m-k))$ |
| Theorem 3.5 | $k > n$ | $\frac{(1+\sqrt{\frac{m}{k}})^2}{(1-\sqrt{\frac{n}{k}})^2}$ | $\frac{(1+\sqrt{\frac{m}{k}})^2}{(1-\sqrt{\frac{n}{k}})^2}$ | $O(mn^2 k)$ |
| Theorem 3.7 $\delta = 1/2$ | $k \geq 32 \cdot n \cdot \ln(4n)$ | $4m$ | $4m$ | $O(mn^2 + k \log k)$ |
| Theorem 3.11 $\delta = 1/2$ | $k = n$ | $(1+\eta)(m-n+1)$ | $(1+\eta)n(m-n+1)$ | $O(mn^3/\log(1+\eta))$ |

Table 1: Summary of various algorithms for Problem 1.1 (our algorithms, as well as previous algorithms). $\mathbf{X} \in \mathbb{R}^{n \times m}$ is a full rank matrix. $k$ denotes the number of sampled columns. $\mathcal{S} \subseteq [m]$ has cardinality $k$. $\delta$ denotes a failure probability, which is assumed to be zero if it is omitted from the description. In the fourth line of the table, $f > 1$ is a parameter which trades accuracy with number of operations. In the fifth line of the table, $\tau$ denotes the coherence of $\mathbf{X}$: $\tau = \frac{m}{n} \max_{i \in [m]} (\mathbf{V}\mathbf{V}^{\mathrm{T}})_{ii}$, where $\mathbf{V} \in \mathbb{R}^{m \times n}$ contains the right singular vectors of $\mathbf{X}$. Lemma 1 in [28] assumes that $\mathbf{X}$ has orthonormal rows; but, this can be extended to arbitrary $\mathbf{X}$ just by applying the result to $\mathbf{V}^{\mathrm{T}}$. In the sixth line of the table, the formulation in [39] assumes that $\mathbf{X}$ is orthonormal, but this can be generalized as well. In the last line of the table, $\eta > 0$ is a parameter which trades accuracy with operations.

| | $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \geq \gamma \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2$; $\gamma =$ | $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2 \geq \gamma \|\mathbf{X}^{\dagger}\|_2^2$; $\gamma =$ |
|---|---|---|
| $k = n$ | $m/n$ | $m$ |
| $k > n$, $k = O(n)$ | $m/k - O(1)$ | $m/k - 1$ |
| $k > n$, $k = \omega(n)$ | $m/k - k/n$ | $m/k - 1$ |

Table 2: Summary of lower bounds for Problem 1.1. By lower bounds, we mean that there is a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ such that for every $\mathcal{S}$, $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^2 \geq \gamma \|\mathbf{X}^{\dagger}\|_{\xi}^2$, for a value of $\gamma$ shown in the table. $\mathcal{S} \subseteq [m]$ has cardinality at most $n \leq k \leq m$. For $\xi = 2$ and $n = k$, the bound is from Lemma 2.2 in [30]; the bound for $\xi = \mathrm{F}$ is an immediate corollary. We prove the other bounds in Section 4.

$O(mn^2 + k \log k)$. For a fixed probability parameter $\delta$ $(0 < \delta < 1)$, and $k \geq \lceil 32n \ln(2n/\delta) \rceil$, the algorithm constructs a subset $\mathcal{S}$ of cardinality at most $k$, such that, for both $\xi = 2, \mathrm{F}$, and with probability $1 - \delta$,

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^2 \leq 4 \cdot m \cdot \|\mathbf{X}^{\dagger}\|_{\xi}^2 .$$

If $\mathbf{X}$ has orthonormal rows, then, the operation count drops to $O(mn + k \log k)$, i.e. linear in the size of the input. The analysis of the algorithm is based on the matrix concentration bound of [47].

Our last algorithm, Algorithm 5 in Theorem 3.11 (Section 3.4) is designed for $k = n$. It is based on the following theoretical contribution (Lemma 3.9 in Section 3.4): if we randomly sample a subset $\mathcal{S}$ of cardinality $k \geq n$ with probability proportional to $\det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}})$, then,

$$\mathbb{E}\left[\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2\right] \leq \frac{m-n+1}{k-n+1} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 \qquad \text{and} \qquad \mathbb{E}\left[\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{2}^2\right] \leq \left(1 + \frac{n(m-k)}{k-n+1}\right) \cdot \|\mathbf{X}^{\dagger}\|_{2}^2.$$

Algorithm 5 finds a subset S of cardinality $k = n$ such that

$$\|\mathbf{X}_{\mathcal{S}}^{-1}\|_2^2 \leq \|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2 \leq (1+\eta) \cdot (m-n+1) \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 \leq (1+\eta) \cdot (m-n+1) \cdot n \cdot \|\mathbf{X}^{\dagger}\|_2^2 ,$$

for any $\eta > 0$ chosen by the user. This bound is deterministic but the bound on the number of operations is probabilistic. Specifically, for any $0 < \delta < 1$, we show that with probability $1 - \delta$, the operation count is $O\left(mn^3 \log \delta^{-1} \log^{-1}(1+\eta)\right)$.

Our volume-sampling-based algorithms for the subset selection problem can be viewed as complementary results to the volume-sampling-based algorithms designed before for low-rank matrix approximation [23]. In low-rank matrix approximation, the subspace spanned by the columns that are selected by volume sampling contains a rank $k$ matrix that approximates the best rank $k$ matrix computed via the SVD; in our case, the objective is different but we show that volume sampling gives useful results as well.

**Lower Bounds.** By lower bounds, we mean that there exists a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ such that for every $\mathcal{S}$ of cardinality $k \geq n$, for $\xi = 2$ or $\xi = \mathrm{F}$, we have $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^2 \geq \gamma \|\mathbf{X}^{\dagger}\|_{\xi}^2$ for some value of $\gamma$ which we call *lower bound*. We develop such lower bounds via, first, relating the subset selection problem to the so-called column-based matrix reconstruction problem [10], and then, employing existing lower bounds [10] for column-based matrix reconstruction. We present these results in Section 4; a summary of lower bounds appears in Table 2. Our lower bounds indicate that some upper bounds of de Hoog and Mattheij [20, 21] as well as ours are the best possible up to constant factors. This resolves an open question in [20, 21].

An alternative way to study optimality is to develop lower bounds of the form $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^2 \geq \gamma \|\mathbf{X}_{\mathcal{S}_{opt}}^{\dagger}\|_{\xi}^2$. However, we were unable to prove such bounds for our algorithms, so we leave this as an interesting open question for future investigation.

**Applications.** In Section 5, we study the connection between low-stretch spanning trees and subset selection. Using a result by Spielman and Woo [50], we prove that the stretch of any tree in an undirected graph equals the Frobenius norm squared of the pseudo-inverse of the sampled matrix that arises by sampling columns from an orthonormal matrix which is a basis for the row space of the so-called node-by-edge incidence matrix of the graph. This incidence matrix contains as many columns as edges in the graph; so, sampling columns from this matrix corresponds to sampling edges from the graph. We then use this reduction to develop novel algorithms for constructing spanning trees with low stretch in undirected graphs. Unfortunately, our algorithms are worse than the available state-of-the-art [26, 1, 41]. We believe, however, that the connection is interesting and might be useful to shed new light on the combinatorial problem of finding a low stretch spanning tree in an undirected graph.

In Section 6 we use the subset selection algorithms of this paper to design novel algorithms for three other problems involving sub-sampling: column-based low-rank matrix reconstruction, sparse solution of least-squares problems, and feature selection in $k$-means clustering.

### 1.1 Related Work

**Greedy Algorithms.** In [20] de Hoog and Mattheij propose the following algorithm for the Frobenius norm version of Problem 1.1. The idea is to proceed by removing one column from $\mathbf{X}$ at a time. In the first iteration of the algorithm, they remove the column with index $i_1$, where

$$i_1 = \arg \min_{i=1,\ldots,m} \mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\right).$$

Let $\mathbf{X}_1 \in \mathbb{R}^{n\times(m-1)}$ be the matrix obtained after removing the $i_1$th column of $\mathbf{X}$. In the second iteration of the algorithm, they remove the column with index $i_2$ such that,

$$i_2 = \arg \min_{i=1,\ldots,m-1} \mathrm{Tr}\left(\left(\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\right),$$

and so on, until $m-k$ columns are removed.

A straightforward implementation of this idea requires $O(mn^3(m-k))$ operations. However, one can use the Sherman-Morrison formula for rank one updates to the inverse of a matrix and improve the operation count to $O(n^3 + mn^2(m-k))$.

Notice that the algorithm just described assumes (implicitly) that in all the iterations, removing a single column does not result in a rank deficient matrix; otherwise, for an iterate $\mathbf{X}_j$ $(j = 1, ..., m-k)$ and a column $\mathbf{x}_i$ $(i = 1, ..., m)$ whose removal will result in a rank deficient matrix, the inverse of $\mathbf{X}_j\mathbf{X}_j^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ is not defined. In [20] it is shown that this algorithm achieves the bound

$$\|\mathbf{X}_{\hat{\mathcal{S}}}^{\dagger}\|_{\mathrm{F}}^2 \leq \frac{m-n+1}{k-n+1} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2.$$

However, the assumption just mentioned is not true in general.

Our algorithms of Theorem 3.1 and Corollary 3.3 use the greedy removal idea as well. However, they find the columns to be removed in a different way. Our algorithms are substantially faster (at least a factor of $n$, and a factor of $n^2$ in some cases) than the algorithm of [20]. Additionally, our algorithms efficiently detect columns whose removal results in a rank deficient matrix, and avoid removing them. So, our algorithms work for any full-rank matrix $\mathbf{X}$, without any restriction.

We also mention that Theorem 1 in [21] describes a similar greedy deterministic algorithm with comparable operation count but slightly worse approximation bounds than the algorithm of [20] (see Table 1 for the precise statement of these results). On the positive side, this algorithm works for any $\mathbf{X}$.

**Rank Revealing Factorizations.** The subset selection problem that we study in this paper has deep connections, which we do not explain in detail, with the so-called Rank-Revealing QR [32] (and also see [12, 17] for a summary of available RRQR algorithms) and Rank-Revealing LU [33] factorizations.

Worth special mention is the seminal work of Gu and Eisenstant [32] on Strong Rank-Revealing QR (RRQR). Algorithm 4 and Theorem 3.2 of [32] are presented for matrices $\mathbf{X} \in \mathbb{R}^{n\times m}$ with $n \geq m$. They can be easily adapted to the case where $n \leq m$ (see Lemma 15 of [9] or Equation 3.1 of [14]). In this case, they can find a subset $\mathcal{S}$ of cardinality $k \leq n$ with some bounds on all the non-zero singular values of $\mathbf{X}_{\mathcal{S}}$. When applied to bound the spectral norm with $k = n$ we have the following bound,

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2 \leq \left(1 + f^2 n\,(m-n)\right) \cdot \|\mathbf{X}^{\dagger}\|_2^2\,.$$

For $f > 1$ and $k = n$, the operation count of this method is $O\left(mn^2 \log_f m\right)$. RRQR approaches can only be used to sample $k \leq n$ columns; extending these approaches to sample arbitrary $k \geq n$ columns, which is the focus of this paper, is not obvious.

**Incoherent Subset Selection.** A recent result by Gittens [28] studies the subset selection problem in the context of the so-called coherence of $\mathbf{X}$. The algorithm uses random sampling of columns. Gittens shows that this simple algorithm gives competitive bounds for incoherent matrices.

**Approximation via Convex Relaxation.** Joshi and Boyd [39] explored the use of convex relaxation to solve Problem 1.1: initially, $\mathbf{X}_{\mathcal{S}}$ is allowed to contain $m$ columns of $\mathbf{X}$, rescaled by weights in the interval $[0, 1]$. This is a convex program, which can be solved, for example, via an interior point algorithm. To get a feasible solution for $\mathbf{X}_{\mathcal{S}}$, one needs an efficient rounding scheme to get strictly 0 or 1 weights. No theoretical results are reported but the method is shown to be efficient in practice.

**Maximum-volume Subsets.** Theorem 1 in de Hoog and Mattheij [20] shows that, for $k \geq n$, if $\mathbf{X}_{\mathcal{S}}$ maximizes $\det\left(\mathbf{X}_{\mathcal{T}}^{\mathrm{T}}\mathbf{X}_{\mathcal{T}}\right)$ among all possible subsets $\mathcal{T}$ of cardinality $k$, then the following two bounds hold, $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2 \leq (1 + n\,(m-k)\,/(k-n+1)) \cdot \|\mathbf{X}^{\dagger}\|_2^2$; and, $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \leq (m-n+1)\,/\,(k-n+1) \cdot n \cdot \|\mathbf{X}^{\dagger}\|_2^2$. Similar spectral norm bounds for $k = n$ were shown before in Eqn. 2.4 of Theorem 2.2 of [36], Lemma 3.4 ($\mu = 1$) in [45], Lemma 2.1 of [30], and Algorithm 3 in [32]. A similar Frobenius norm bound for $k = n$ was shown before in Eqn. 2.13 of Theorem 2.3 of [36]. Notice that all these results do not imply any algorithm other than the naive procedure of testing all the $\binom{m}{k}$ possible subsets of cardinality $k$ (this procedure has an exponential operation count).

Our Lemma 3.9 proves a similar result, which states that if one samples $\mathcal{S}$ with probability proportional to $\det\left(\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\mathbf{X}_{\mathcal{S}}\right)$, then, the same bounds hold in expectancy. Now, recent polynomial-time implementations ($O(mn^3)$ operations) of such determinant-based random sampling [22, 34] allow us to design efficient algorithms ($O(mn^3)$ operations with high probability) that achieve only slightly larger bounds.

Finally, note that the strong RRQR algorithm of [32] can also find a local maximum-volume subset. By local maximum-volume subset, we mean that the volume of the subset found is always bigger than the volume of any subset obtained by interchanging a single column.

**Computational Complexity of Subset Selection.** In [19], Civril and Magdon-Ismail study the spectral norm version of Problem 1.1, as well as three other similar subset selection problems, from a complexity theory point of view. They show that these problems are NP-hard. They give special emphasis to the problem of finding a subset $\mathcal{S}$ for which $\mathbf{X}_{\mathcal{S}}$ has maximum volume, i.e. $\det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\top})$ is maximized. As we discussed above, the problem of finding the subset with maximum volume is connected to Problem 1.1.

The computational complexity of finding a maximum volume subset was also investigated in the computational geometry literature. The problem is stated differently: finding a large simplex in a V-polytope. NP-hardness was established in [44], and exponential inapproximability was established in [40].

**Variants of the Subset Selection Problem.** Other variants of subset selection have been studied extensively in numerical linear algebra and computer science. Most of this work focused on spectral norm and the case of $\mathbf{X}_{\mathcal{S}}$ containing *rescaled* columns from $\mathbf{X}$ (Theorem 3.1 in [47]; Theorem 11 in [57]; Theorem 3.1 in [4]) or $\mathbf{X}_{\mathcal{S}}$ containing *linear combinations* of columns from $\mathbf{X}$ (Lemma 3.15 of [43]; Lemma 6 of [48]; Theorem 1.3 of [53]). We should note that all these results give much better approximation bounds than our bounds for the spectral norm version of Problem 1.1. For example, the deterministic algorithm in Theorem 3.1 in [4], for any $\epsilon > 0$, selects and appropriately rescales $O(n/\epsilon^2)$ columns from $\mathbf{X}$ and guarantees an approximation bound $1 + \epsilon$.

We believe that the main reason of why our bounds are worse than the bounds in these two variants of the problem is the absence of rescaling. Selecting actual columns from $\mathbf{X}$ keeps less information from the original matrix than selecting rescaled columns or keeping linear combinations of the columns. In our case we keep $nk$ numbers to summarize $\mathbf{X}$. However, when one considers rescaling, she keeps $nk + k$

numbers, i.e. $k$ columns of $\mathbf{X}$ and a weight for each column. Taking linear combinations of the columns of $\mathbf{X}$ essentially means multiplying $\mathbf{X}$ with some $m \times r$ (e.g. $r = O(n)$) matrix $\mathbf{W}$, thus the information used to represent $\mathbf{X}_{\mathcal{S}}$ is $nk + O(mn)$. Our lower bounds formally argue that selecting actual columns from $\mathbf{X}$ can not give as tight bounds as the bounds obtained from the two variants of the problem mentioned in this paragraph.

Finally, we mention that all these algorithms have found many applications in numerous problems involving subsampling: least-squares regression [3]; column-based low-rank matrix approximation [12]; spectral graph sparsification [52]; and, dimensionality reduction in clustering [13].

**Restricted Invertibility.** Bourgain and Tzafriri restricted invertibility result [7] states that there exists a universal constant $C$ such that for every square invertible matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ whose columns have unit $\ell_2$ norm, one can find a subset $\mathcal{S} \subseteq [n]$ of cardinality at least $Cn/\|\mathbf{A}\|_2^2$ such that

$$\|\mathbf{A}_{\mathcal{S}}\|_2 \cdot \|\mathbf{A}_{\mathcal{S}}^{\dagger}\|_2 \leq \sqrt{3}\,.$$

Given $\mathbf{A}$, finding such a subset $\mathcal{S}$ is another variant of column subset selection. Tropp gave the first polynomial (randomized) algorithm for restricted invertibility [54]. A deterministic algorithm was recently suggested by Spielman and Srivastava [49]. However, restricted invertibility deals with selecting fewer than $n$ columns that maximize the smallest non-zero singular value, while Problem 1.1 deals with selecting at least $n$ columns so that the matrix is full-rank, and the smallest singular value is as large as possible. So, the problems are similar, but different.

## 2 Preliminaries

**Basic Notation.** We use $[n]$ to denote the set $\{1, \ldots, n\}$. We use $\mathbf{X}, \mathbf{Y} \ldots$ to denote matrices; $\mathbf{x}, \mathbf{y} \ldots$ to denote column vectors. We denote the columns of $\mathbf{X} \in \mathbb{R}^{n \times m}$ by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathbb{R}^n$; $\mathbf{x}_i$ is column $i$ of $\mathbf{X}$. $\mathbf{I}_m$ is the $m \times m$ identity matrix; $\mathbf{0}_{n \times m}$ is the $n \times m$ matrix of zeros; $\boldsymbol{e}_i$ is the $i$th standard basis vector (whose dimensionality will be clear from the context): all entries are zero except the $i$th entry which equals one. $\mathbf{X}_{ij}$ or $(\mathbf{X})_{ij}$ denotes the $(i,j)$th element of $\mathbf{X}$. $\mathbf{v}_{ij}$ denotes the $j$th element of a vector $\mathbf{v}_i$. Logarithms are base two. We abbreviate "independent identically distributed" to "i.i.d". Finally, for a set $A$, we denote by $C(A,k)$ the set of all subsets of $A$ of cardinality $k$.

**Sampling Columns.** In the context of Problem 1.1, $\mathcal{S}$ is a set of cardinality $1 < k \leq m$, which contains some subset of the natural numbers from $1, 2, ..., m$ (repetition of numbers is *not* allowed). $\mathbf{X}_{\mathcal{S}}$ contains the columns of some matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, indicated in $\mathcal{S}$; sometimes we will use $(\mathbf{X})_{\mathcal{S}}$ to denote the same matrix. The columns of $\mathbf{X}_{\mathcal{S}}$ are ordered consistently with their order in $\mathbf{X}$: if $i, j$ are elements from $\mathcal{S}$ and $i < j$, then, the $i$th column of $\mathbf{X}$ will appear before the $j$th column of $\mathbf{X}$ in $\mathbf{X}_{\mathcal{S}}$. Finally, $\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}$ means $(\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}$ and $\mathbf{X}_{\mathcal{S}}^{\dagger}$ means $(\mathbf{X}_{\mathcal{S}})^{\dagger}$.

**Singular Value Decomposition.** The Singular Value Decomposition (SVD) of $\mathbf{X} \in \mathbb{R}^{n \times m}$ is:

$$\mathbf{X} = \underbrace{\begin{pmatrix} \mathbf{U}_r & \mathbf{U}_{\rho-r} \end{pmatrix}}_{\mathbf{U} \in \mathbb{R}^{n \times \rho}} \underbrace{\begin{pmatrix} \boldsymbol{\Sigma}_r & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\rho-r} \end{pmatrix}}_{\boldsymbol{\Sigma} \in \mathbb{R}^{\rho \times \rho}} \underbrace{\begin{pmatrix} \mathbf{V}_r^{\mathrm{T}} \\ \mathbf{V}_{\rho-r}^{\mathrm{T}} \end{pmatrix}}_{\mathbf{V}^{\mathrm{T}} \in \mathbb{R}^{\rho \times m}},$$

with singular values $\sigma_1 \geq \ldots \sigma_r \geq \sigma_{r+1} \geq \ldots \geq \sigma_\rho > 0$. Here, $\rho = \mathrm{rank}(\mathbf{X})$ and $r$ is some rank parameter $1 \leq r \leq \rho$. We will often denote $\sigma_1$ as $\sigma_{\max}$ and $\sigma_\rho$ as $\sigma_{\min}$, and will use $\sigma_i(\mathbf{X})$ to denote the $i$-th singular value of $\mathbf{X}$ when the matrix is not clear from the context. The matrices $\mathbf{U}_r \in \mathbb{R}^{n \times k}$ and $\mathbf{U}_{\rho-r} \in \mathbb{R}^{n \times (\rho-r)}$

contain the left singular vectors of $\mathbf{X}$; and, similarly, the matrices $\mathbf{V}_r \in \mathbb{R}^{m\times k}$ and $\mathbf{V}_{\rho-r} \in \mathbb{R}^{m\times(\rho-r)}$ contain the right singular vectors of $\mathbf{X}$. Finally, we repeatedly use the following column representation for the matrix $\mathbf{V}$: $\mathbf{V}^{\mathrm{T}} = \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m]$. Here, the $\mathbf{y}_i$'s are vectors in $\mathbb{R}^n$.

**Moore-Penrose Pseudo-inverse.** Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ with SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$. Then, $\mathbf{X}^\dagger = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}} \in \mathbb{R}^{m\times n}$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{X}$ ($\mathbf{\Sigma}^{-1}$ is the inverse of $\mathbf{\Sigma}$).

**Lemma 2.1** (Fact 6.4.12 in [5])**.** *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times \ell}$*, and assume that* $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{B}) = n$*. Then,* $(\mathbf{AB})^\dagger = \mathbf{B}^\dagger\mathbf{A}^\dagger$*.*

**Lemma 2.2.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times m}$ *be a full rank matrix with* $m \geq n$*. Let* $\mathbf{B}$ *be an invertible* $m \times m$ *matrix. Then*

$$\|(\mathbf{AB})^\dagger\|_2 \leq \|\mathbf{A}^\dagger\|_2 \cdot \|\mathbf{B}^{-1}\|_2 .$$

*Proof.*

$$\begin{aligned}
\|(\mathbf{AB})^\dagger\|_2 = (\sigma_{\min}(\mathbf{AB}))^{-1} = (\sigma_{\min}(\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}))^{-1} &= \left(\min_{\mathbf{x}\neq 0} \frac{\|\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}\right)^{-1} \\
&\leq \left(\min_{\mathbf{x}\neq 0} \frac{\|\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2}{\|\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2} \cdot \frac{\|\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}\right)^{-1} \\
&\overset{(*)}{\leq} \left(\min_{\mathbf{x}\neq 0} \frac{\|\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2}{\|\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2} \cdot \min_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}\right)^{-1} \\
&\leq \left(\min_{\mathbf{y}\neq 0} \frac{\|\mathbf{B}^{\mathrm{T}}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \cdot \min_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}\right)^{-1} \\
&= (\sigma_{\min}(\mathbf{B}^{\mathrm{T}}) \cdot \sigma_{\min}(\mathbf{A}^{\mathrm{T}}))^{-1} \\
&= \|\mathbf{A}^\dagger\|_2 \cdot \|\mathbf{B}^{-1}\|_2
\end{aligned}$$

In $(*)$ we use the fact that $\mathbf{A}^{\mathrm{T}}$ is a full rank matrix with more rows than columns (so $\|\mathbf{A}^{\mathrm{T}}\mathbf{x}\|_2 \neq 0$ for $\mathbf{x} \neq 0$). ∎

**Column Exchanges and Cramer's rule.** For a matrix $\mathbf{A}$, an index $i$ and a vector $\mathbf{v}$, we denote by $\mathbf{A}(i \to \mathbf{v})$ the matrix obtained after replacing the $i$th column of $\mathbf{A}$ by $\mathbf{v}$. Notice that for square matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension we have $\det(\mathbf{A})\det(\mathbf{B}(i \to \mathbf{v})) = \det((\mathbf{AB})(i \to \mathbf{Av}))$. For an invertible square matrix $\mathbf{A}$, recall Cramer's rule, which gives a formula for computing the components of $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ in terms of determinants. In our notation, the rule states that $\mathbf{x}_i$, the $i$th position in $\mathbf{x}$, is

$$\mathbf{x}_i = \frac{\det(\mathbf{A}(i \to \mathbf{b}))}{\det(\mathbf{A})} .$$

**Volume Sampling.** Let $\mathbf{X}$ be a full rank matrix of dimensions $n \times m$ with $m \geq n$, and let $n \leq k \leq m$ be some integer. Given a subset $\mathcal{S} \in C([m], k)$ define the probability of $\mathcal{S}$ by

$$P_{\mathcal{S}} = \frac{\det\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)}{\sum_{\mathcal{T}\in C([m],k)} \det\left(\mathbf{X}_{\mathcal{T}}\mathbf{X}_{\mathcal{T}}^{\mathrm{T}}\right)} .$$

The values $\{P_{\mathcal{S}}\}_{\mathcal{S}\in C([m],k)}$ define a distribution over the sets in $C([m], k)$. We denote this distribution by $\mathrm{VolSamp}(\mathbf{X}, k)$. That is, we write $\mathcal{S} \sim \mathrm{VolSamp}(\mathbf{X}, k)$ to denote that $\mathcal{S}$ is a random subset which assumes value in $C([m], k)$, whose distribution is defined by

$$\Pr(\mathcal{S} = \mathcal{T}) = P_{\mathcal{T}} .$$

We call this sampling distribution *volume sampling* due to the fact that $\det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}})^{1/2}$ is the volume of the parallelpiped defined by the rows of $\mathbf{X}_{\mathcal{S}}$. Notice that if $\mathbf{A}$ is a square non-singular matrix then $\mathrm{VolSamp}(\mathbf{AX}, k) = \mathrm{VolSamp}(\mathbf{X}, k)$ (this follows from the fact that $\det((\mathbf{AX})_{\mathcal{S}}(\mathbf{AX})_{\mathcal{S}}^{\mathrm{T}}) = \det(\mathbf{A})^2 \det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}})$, for every $\mathcal{S}$).

An efficient algorithm for sampling a set from $\mathrm{VolSamp}(\mathbf{X}, n)$ was first suggested by Deshpande and Rademacher [22]. This algorithm was recently improved by Guruswami and Sinop, who showed how to sample such a subset with $O(n^3 m)$ operations [34]. There is currently no algorithm for sampling from $\mathrm{VolSamp}(\mathbf{X}, k)$ for an arbitrary $k \geq n$.

**Other Known Results.** In addition we use the following two known results.

**Lemma 2.3** (Special case of the Cauchy-Binet formula)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{B} \in \mathbb{R}^{n\times m}$, and $m \geq n$. Then,*

$$\det(\mathbf{AB}^{\mathrm{T}}) = \sum_{\mathcal{S}\in C([m],n)} \det(\mathbf{A}_{\mathcal{S}}) \det(\mathbf{B}_{\mathcal{S}}^{\mathrm{T}}).$$

**Lemma 2.4** (Theorem 1.2.12 in [37])**.** *Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$, denote the eigenvalues of $\mathbf{A} \in \mathbb{R}^{m\times m}$. Let $1 \leq k \leq m$. Then,*

$$\sum_{\mathcal{S}\in C([m],k)} \det\left(\mathbf{A}_{\mathcal{S},\mathcal{S}}\right) = \sum_{\mathcal{S}\in C([m],k)} \prod_{i\in\mathcal{S}} \lambda_i.$$

*Here, $\mathbf{A}_{\mathcal{S},\mathcal{S}} \in \mathbb{R}^{k\times k}$ denotes the submatrix of $\mathbf{A}$ corresponding to the rows and the columns in $\mathcal{S} \subseteq [m]$, which has cardinality $k$.*

Lemma 2.4 expresses the $k$-th elementary symmetric function of the eigenvalues of $\mathbf{A}$ as the sum of the $k$-by-$k$ principal minors of $\mathbf{A}$.

# 3 Algorithms

## 3.1 Deterministic Greedy Removal

The algorithm of this section uses the same greedy removal strategy as in [20], but it is faster, since it exploits the SVD decomposition of the matrix and the ability to quickly update it. Additionally, our algorithm efficiently detects columns whose removal results in a rank deficient matrix, and avoids removing them (see the discussion in Section 1.1). We prove that our algorithm achieves the same approximation bounds as in [20]. The proof of [20] does not apply to our algorithm, since [20] assumes implicitly that in all the iterations, removing a single column does not result in a rank deficient matrix.

**Theorem 3.1.** *Fix $\mathbf{X} \in \mathbb{R}^{n\times m}$ ($m > n$, $\mathrm{rank}(\mathbf{X}) = n$) and sampling parameter $m \geq k \geq n$. Algorithm 1 needs $O\left(mn^2 + mn\left(m-k\right)\right)$ operations and deterministically constructs a set $\mathcal{S} \subseteq [m]$ of cardinality $k$ with*

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \leq \frac{m-n+1}{k-n+1} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 \qquad \text{and} \qquad \|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2 \leq \frac{m-n+1}{k-n+1} \cdot n \cdot \|\mathbf{X}^{\dagger}\|_2^2.$$

*Moreover, if $\mathbf{X}$ contains orthonormal rows, the operation count is $O\left(mn\left(m-k\right)\right)$.*

Before proceeding with the proof we state an auxiliary lemma. However, we defer the proof to Section 3.4 since this Lemma is a corollary of a Theorem that appears in that section.

**Lemma 3.2.** *Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ ($m \geq n$) be a full rank matrix. There exists a subset $\mathcal{S} \subset [m]$ of cardinality $m-1$ such that $\mathbf{X}_{\mathcal{S}}$ is full rank and $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \leq \frac{m-n+1}{m-n} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2$.*

**Input:** $\mathbf{X} \in \mathbb{R}^{n\times m}$, sampling parameter $n \leq k \leq m$.
**Output:** Set $\mathcal{S} \subseteq [m]$ of cardinality $k$.

1: $\mathcal{S}_0 \leftarrow [m]$
2: Compute the SVD of $\mathbf{X}_{\mathcal{S}_0}$: $\mathbf{X}_{\mathcal{S}_0} = \mathbf{U}^{(0)}\mathbf{\Sigma}^{(0)}\mathbf{Y}^{(0)}$
3: **for** $i = 1, 2, \ldots, m-k$ **do**
4: Let the singular values of $\mathbf{X}_{\mathcal{S}_{i-1}}$ be $\sigma_1^{(i-1)}, \ldots, \sigma_n^{(i-1)}$.
5: Let the columns of $\mathbf{Y}^{(i-1)}$ be $\{\mathbf{y}_r^{(i-1)}\}_{r\in\mathcal{S}_{i-1}}$.
Denote by $\mathbf{y}_{lr}^{(i-1)}$ the $l$-th element of $\mathbf{y}_r^{(i-1)}$.
6: $j_i \leftarrow \arg\min_{r\in\mathcal{S}_{i-1};\|\mathbf{y}_r^{(i-1)}\|_2<1} \left( \frac{\sum_{l=1}^{n}\left(\mathbf{y}_{lr}^{(i-1)}/\sigma_l^{(i-1)}\right)^2}{1-\|\mathbf{y}_r^{(i-1)}\|_2^2} \right)$.
{See proof on how to implement this step in a stable manner.}
7: $\mathcal{S}_i \leftarrow \mathcal{S}_{i-1} - \{j_i\}$
8: Downdate the SVD of $\mathbf{X}_{\mathcal{S}_{i-1}}$ to obtain an SVD of $\mathbf{X}_{\mathcal{S}_i} = \mathbf{U}^{(i)}\mathbf{\Sigma}^{(i)}\mathbf{Y}^{(i)}$.
{Using an algorithm described in [31]}
9: **end for**
10: **return** $\mathcal{S}$

**Algorithm 1**: A deterministic greedy removal algorithm for subset selection (Theorem 3.1).

*Proof of Theorem 3.1.* The spectral norm bound is immediate from the Frobenius norm bound using the fact that for any matrix $\mathbf{B}$, $\|\mathbf{B}\|_2^2 \leq \|\mathbf{B}\|_{\mathrm{F}}^2 \leq \mathrm{rank}(\mathbf{B}) \cdot \|\mathbf{B}\|_2^2$. So, we prove only the Frobenius norm bound.

Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{\Sigma}\mathbf{Y}$ be an SVD of $\mathbf{X}$ (see also Section 2 for useful notation). First, we argue that $\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ is singular if and only if $\|\mathbf{y}_i\|_2 = 1$. Suppose $\mathrm{rank}(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}) < n$. Notice that, $\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \mathbf{U}\mathbf{\Sigma}\left(\mathbf{I}_n - \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\right)\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}$. The matrix $\mathbf{U}$ is full rank (it is square unitary), so we find that $\mathbf{\Sigma}\left(\mathbf{I}_n - \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\right)\mathbf{\Sigma}$ is singular. We now observe that $\mathbf{\Sigma}$ is full rank (it is diagonal with positive values on the diagonal) as well, so we find that $\mathbf{I}_n - \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}$ is singular. That can hold only if $\|\mathbf{y}_i\|_2 = 1$. Therefore, comparing the norm of $\mathbf{y}_i$ with 1 is an efficient way (once we have an SVD) under exact arithmetic to detect if $\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ is singular.

We proceed with some calculations. Let $\sigma_1, \sigma_2, \ldots, \sigma_n$ denote the singular values of $\mathbf{X}$. Fix an index $i$. If $\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ is not singular, then,

$$\begin{aligned}
\mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\right) &\overset{(a)}{=} \mathrm{Tr}\left(\left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^{\mathrm{T}} - \mathbf{U}\mathbf{\Sigma}\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}\right)^{-1}\right) \\
&\overset{(b)}{=} \mathrm{Tr}\left(\mathbf{\Sigma}^{-2} + \frac{\mathbf{\Sigma}^{-1}\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\mathbf{\Sigma}^{-1}}{1-\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i}\right) \\
&\overset{(c)}{=} \mathrm{Tr}\left(\mathbf{\Sigma}^{-2}\right) + \mathrm{Tr}\left(\frac{\mathbf{\Sigma}^{-1}\mathbf{y}_i\mathbf{y}_i^{\mathrm{T}}\mathbf{\Sigma}^{-1}}{1-\mathbf{y}_i^{\mathrm{T}}\mathbf{y}_i}\right) \\
&\overset{(d)}{=} \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 + \frac{1}{1-\|\mathbf{y}_i\|_2^2} \cdot \mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\mathbf{y}_i\left(\mathbf{\Sigma}^{-1}\mathbf{y}_i\right)^{\mathrm{T}}\right) \\
&\overset{(e)}{=} \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 + \frac{\|\mathbf{\Sigma}^{-1}\mathbf{y}_i\|_2^2}{1-\|\mathbf{y}_i\|_2^2} \\
&\overset{(f)}{=} \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 + \frac{\sum_{j=1}^{n}\left(\mathbf{y}_{ji}/\sigma_j\right)^2}{1-\|\mathbf{y}_i\|_2^2}
\end{aligned}$$

$(a)$ follows by replacing the SVD of $\mathbf{X}$. $(b)$ follows from the Sherman-Morrison formula (recall that we

assume that $\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ is not singular, so $\|\mathbf{y}_i\|_2 \neq 1$) and the identity $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_n$ , which also implies that $\mathrm{Tr}(\mathbf{U}\mathbf{A}\mathbf{U}^{\mathrm{T}}) = \mathrm{Tr}(\mathbf{A})$. $(c)$ follows by the linearity of the trace operator. $(d)$ follows from the fact that $1/(1 - \|\mathbf{y}_i\|_2^2)$ is a scalar. $(e)$ follows from the fact that for any matrix $\mathbf{B}$ we have $\mathrm{Tr}\left(\mathbf{B}\mathbf{B}^{\mathrm{T}}\right) = \|\mathbf{B}\|_{\mathrm{F}}^2$; in our case, we apply this equality to $\mathbf{B} = \mathbf{\Sigma}^{-1}\mathbf{y}_i$. $(f)$ follows because $\mathbf{\Sigma}$ is diagonal.

**Algorithm.** To facilitate the description of the algorithm, we assume the columns in $\mathbf{X}_{\mathcal{S}}$ are indexed using their index in $\mathbf{X}$. Our algorithm constructs $\mathcal{S}$ by iteratively removing columns. That is, we start with a complete subset $\mathcal{S}_0 = [m]$. Then we proceed with $m - k$ iterations, since the goal is to select $k$ columns. We reserve the index $i$ to refer to the iterations of the algorithm; so, $i = 1, 2, \ldots, m$.

Each iteration $i$ of the algorithm starts with some $\mathcal{S}_{i-1} \subseteq [m]$ of cardinality $m - i + 1$, and removes one index from it, to obtain $\mathcal{S}_i \subset \mathcal{S}_{i-1}$ of cardinality $m - i$. Our algorithm does exactly $m - k$ iterations, hence it returns $\mathcal{S}_{m-k}$ of cardinality $k$. Additionally, in each iteration we maintain an SVD of the current subset, $\mathbf{X}_{\mathcal{S}_i} = \mathbf{U}^{(i)}\mathbf{\Sigma}^{(i)}\mathbf{Y}^{(i)}$. We denote the singular values of $\mathbf{X}_{\mathcal{S}_i}$ by $\sigma_1^{(i)}, \ldots, \sigma_n^{(i)}$, and the columns of $\mathbf{Y}^{(i)}$ by $\{\mathbf{y}_r^{(i)}\}_{r\in\mathcal{S}_i}$.

Our algorithm begins by computing the SVD of $\mathbf{X}_{\mathcal{S}_0} = \mathbf{X}$. Then, for $i = 1, 2, \ldots, m - k$, iteration $i$ has two stages:

1. Finding an index $j_i$ to remove. We then set $\mathcal{S}_i = \mathcal{S}_{i-1} - \{j_i\}$.
2. Updating the SVD $\mathbf{X}_{\mathcal{S}_i} = \mathbf{U}^{(i)}\mathbf{\Sigma}^{(i)}\mathbf{Y}^{(i)}$. The algorithm needs only $\mathbf{\Sigma}^{(i)}$ and $\mathbf{Y}^{(i)}$ (no need to downdate $\mathbf{U}^{(i)}$).

We now describe each stage in detail. Our algorithm implements the greedy removal idea of Theorem 2 in [20], so $j_i$ is selected as to minimize $\|\mathbf{X}_{\mathcal{S}_i}^{\dagger}\|_{\mathrm{F}}^2$ (subject to constraint that $\mathcal{S}_i$ is a subset of $\mathcal{S}_{i-1}$). Specifically, the formula for $j_i$ is

$$j_i = \arg \min_{r\in\mathcal{S}_{i-1};\|\mathbf{y}_r^{(i-1)}\|_2<1} \left( \frac{\sum_{l=1}^n \left(\mathbf{y}_{lr}^{(i-1)}/\sigma_l^{(i-1)}\right)^2}{1 - \|\mathbf{y}_r^{(i-1)}\|_2^2} \right) .$$

In the last equation, $\mathbf{y}_{lr}^{(i-1)}$ is the $(l, r)$ element of $\mathbf{Y}^{(i-1)}$ or, equivalently, the $l$ element of $\mathbf{y}_r^{(i-1)}$. Equation (1) guarantees that $j_i$ minimizes $\|\mathbf{X}_{\mathcal{S}_i}^{\dagger}\|_{\mathrm{F}}^2$.

As for the second stage, we simply downdate the SVD of $\mathbf{X}_{\mathcal{S}_{i-1}}$ to obtain an SVD of $\mathbf{X}_{\mathcal{S}_i}$, using the algorithm described in [31].

**Stability issues.** Equation (3.1) is unstable since it can potentially suffer from catastrophic cancellations when $\|\mathbf{y}_r^{(i-1)}\|_2 \approx 1$. However, to find the minimizer we need to do only comparisons. That is, we need to be able to determine for two indices $g, h \in \mathcal{S}_{i-1}$ whether

$$\frac{\sum_{l=1}^n \left(\mathbf{y}_{lg}^{(i-1)}/\sigma_l^{(i-1)}\right)^2}{1 - \|\mathbf{y}_g^{(i-1)}\|_2^2} \leq \frac{\sum_{l=1}^n \left(\mathbf{y}_{lh}^{(i-1)}/\sigma_l^{(i-1)}\right)^2}{1 - \|\mathbf{y}_h^{(i-1)}\|_2^2}$$

or not. It is easy to verify that provided $\|\mathbf{y}_g^{(i-1)}\|_2 < 1$ and $\|\mathbf{y}_h^{(i-1)}\|_2 < 1$, the last equation holds if and only if

$$\sum_{l=1}^n \left(\mathbf{y}_{lg}^{(i-1)}/\sigma_l^{(i-1)}\right)^2 + \|\mathbf{y}_g^{(i-1)}\|_2^2 \cdot \sum_{l=1}^n \left(\mathbf{y}_{lh}^{(i-1)}/\sigma_l^{(i-1)}\right)^2 \leq \sum_{l=1}^n \left(\mathbf{y}_{lh}^{(i-1)}/\sigma_l^{(i-1)}\right)^2 + \|\mathbf{y}_h^{(i-1)}\|_2^2 \cdot \sum_{l=1}^n \left(\mathbf{y}_{lg}^{(i-1)}/\sigma_l^{(i-1)}\right)^2 .$$

**Input:** $\mathbf{X} \in \mathbb{R}^{n\times m}$, sampling parameter $n \le k \le m$.
**Output:** Set $\mathcal{S} \subseteq [m]$ of cardinality $k$.

1: Compute the matrix $\mathbf{V} \in \mathbb{R}^{m\times n}$ of the top $n$ right singular vectors of $\mathbf{X}$.
2: Run Algorithm 1 with inputs $\mathbf{V}^{\mathrm{T}}$ and $k$ to obtain $\mathcal{S}$ of cardinality $k$.
3: **return** $\mathcal{S}$

**Algorithm 2**: A deterministic greedy removal algorithm for subset selection (Corollary 3.3).

The last equation does not do any substraction, so it does not suffer from catastrophic cancellations.

Another issue with equation (3.1) is that an index $h \in \mathcal{S}_{i-1}$ is a candidate minimizer only if $\|\mathbf{y}_h^{(i-1)}\|_2 < 1$. Under inexact arithmetic that will always be the case, even if removing the column results in a rank deficient system. This issue can be solved by replacing the test $\|\mathbf{y}_h^{(i-1)}\|_2 < 1$ with $\|\mathbf{y}_h^{(i-1)}\|_2 < 1-\tau$ for some small threshold $\tau$.

**Approximation bound.** Recall that the spectral norm bound follows immediately from the Frobenius norm bound. We now focus on proving the Frobenius norm bound.

We will show using induction that

$$\|\mathbf{X}_{\mathcal{S}_i}^{\dagger}\|_{\mathrm{F}}^2 \le \frac{m-n+1}{m-i-n+1} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 . \tag{1}$$

Since our algorithm returns $\mathcal{S}_{m-k}$ the claim follows from (1).

Equation (1) trivially holds for $i = 0$. Assume it holds for $i-1$. We now show it holds for $i+$. Note that the cardinality of $\mathcal{S}_{i-1}$ is $m-i+1$. Lemma 3.2 ensures that there exists a subset $\mathcal{T}_i \subset \mathcal{S}_{i-1}$ of cardinality $m-i$ such that

$$\|\mathbf{X}_{\mathcal{T}_i}^{\dagger}\|_{\mathrm{F}}^2 \le \frac{m-i-n+2}{m-i-n+1} \cdot \|\mathbf{X}_{\mathcal{S}_i}^{\dagger}\|_{\mathrm{F}}^2 \le \frac{m-i-n+2}{m-i-n+1} \cdot \frac{m-n+1}{m-i-n+2} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 = \frac{m-n+1}{m-i-n+1} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 .$$

Our algorithm finds a subset $\mathcal{S}_i \subset \mathcal{S}_{i-1}$ of cardinality $m-i$ with minimal $\|\mathbf{X}_{\mathcal{S}_i}^{\dagger}\|_{\mathrm{F}}^2$, so

$$\|\mathbf{X}_{\mathcal{S}_i}^{\dagger}\|_{\mathrm{F}}^2 \le \|\mathbf{X}_{\mathcal{T}_i^{\dagger}}\|_{\mathrm{F}}^2 \le \frac{m-n+1}{m-i-n+1} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2 .$$

**Operation count.** First, one needs $O(mn^2)$ operations to compute an SVD of $\mathbf{X}$. Now, at iteration $i$, computing $j_i$ requires $O((m-i+1)n)$ operations. Downdating the SVD to find $\mathbf{\Sigma}^{(i)}$ and $\mathbf{Y}^{(i)}$ can be done be done in $O((m-i+1)n)$ operations [1]. There are $m-k$ iterations, so overall, $O\left(mn^2 + mn\,(m-k)\right)$ operations suffice. If $\mathbf{X}$ has orthonormal rows, the operation count is just $O\left(mn\,(m-k)\right)$ because the initial SVD is available. ∎

We now describe an algorithm which achieves a slightly worse Frobenius norm bound than the bound in the previous theorem but a slightly better spectral norm bound. The set $\mathcal{S}$ of the corollary below and the one in Theorem 3.1 might be different.

[1] This is precisely Problem 3 in page 794 of [31]; the third paragraph in page 795 of [31] argues that this problem can be solved in $O(mn\log^2\epsilon)$ operations, where $\epsilon$ is the machine precision. In our analysis we ignore the $\log^2\epsilon$ term since $\epsilon$ is constant, and $\log^2\epsilon$ is not too big since typically $\epsilon \approx 10^{-16}$. Ignoring such terms is common in the analysis of SVD-type algorithms.

**Corollary 3.3.** *Fix $\mathbf{X} \in \mathbb{R}^{n\times m}$ ($m > n$, $\mathrm{rank}(\mathbf{X}) = n$) and sampling parameter $k \geq n$. Algorithm 2 needs $O\left(mn^2 + mn\,(m-k)\right)$ operations and deterministically constructs a set $\mathcal{S} \subseteq [m]$ of cardinality $k$ with*

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \leq \frac{m-n+1}{k-n+1} \cdot n \cdot \|\mathbf{X}^{\dagger}\|_2^2.$$

*Also, for $i = 1, \ldots, n$ we have*

$$\sigma_i^2(\mathbf{X}) \cdot \left(1 + \frac{n\,(m-k)}{k-n+1}\right)^{-1} \leq \sigma_i^2(\mathbf{X}_{\mathcal{S}})\,.$$

*In particular,*

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2 \leq \left(1 + \frac{n\,(m-k)}{k-n+1}\right) \cdot \|\mathbf{X}^{\dagger}\|_2^2\,.$$

*Moreover, if $\mathbf{X}$ contains orthonormal rows, the operation count is $O\left(mn\,(m-k)\right)$.*

*Proof.* Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be an SVD of $\mathbf{X}$ with $\mathbf{U} \in \mathbb{R}^{m\times m}$, $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$, and $\mathbf{V} \in \mathbb{R}^{n\times m}$. The algorithm of this corollary consists of applying the algorithm of Theorem 3.1 to the matrix $\mathbf{V}^{\mathrm{T}}$ from the SVD of $\mathbf{X}$. Let $\mathcal{S} \subseteq [m]$ be the set found by the algorithm, and let $\bar{\mathcal{S}} = [m] - \mathcal{S}$. Corollary 2 of [20] asserts that $\|\mathbf{X}_{\mathcal{S}}^{\dagger} \cdot \mathbf{X}_{\bar{\mathcal{S}}}\|_{\mathrm{F}}^2 \leq \frac{n(m-k)}{k-n+1}$. Now, Corollary 1 in [20] indicates that, if such a bound holds for $\|\mathbf{X}_{\mathcal{S}}^{\dagger} \cdot \mathbf{X}_{\bar{\mathcal{S}}}\|_{\mathrm{F}}^2$, then,

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \leq \frac{m-n+1}{k-n+1} \cdot n \cdot \|\mathbf{X}^{\dagger}\|_2^2; \qquad \text{for } i = 1, ..., n : \sigma_i^2(\mathbf{X}) \cdot \left(1 + \frac{n\,(m-k)}{k-n+1}\right)^{-1} \leq \sigma_i^2(\mathbf{X}_{\mathcal{S}})\,.$$

■

### 3.2 Deterministic Greedy Selection

The algorithm of this section is an application of the deterministic algorithm presented in [10], which is, in turn, a generalization of an algorithm from [4]. In particular, we use Lemma 10 from [10].

**Lemma 3.4** (Dual Set Spectral Sparsification, Lemma 10 in [10].)**.** *Let $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ and $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ be two equal cardinality decompositions of identity matrices: $\mathbf{v}_i \in \mathbb{R}^n$ ($n < m$), $\mathbf{u}_i \in \mathbb{R}^{\ell}$ ($\ell \leq m$), $\sum_{i=1}^m \mathbf{v}_i\mathbf{v}_i^{\mathrm{T}} = \mathbf{I}_n$, and $\sum_{i=1}^m \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} = \mathbf{I}_{\ell}$. Given an integer $k$ with $n < k \leq m$, there exists an algorithm that computes a set of weights $s_i \geq 0$ ($i = 1, \ldots, m$) at most $k$ of which are non-zero, such that*

$$\sigma_n\left(\sum_{i=1}^m s_i\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\right) \geq \left(1 - \sqrt{\frac{n}{k}}\right)^2 \qquad \text{and} \qquad \sigma_1\left(\sum_{i=1}^m s_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\right) \leq \left(1 + \sqrt{\frac{\ell}{k}}\right)^2.$$

*The algorithm is deterministic and needs at most $O\left(km\left(n^2 + \ell^2\right)\right)$ operations. Moreover, if the set $\mathcal{U}$ contains vectors from the standard basis from $\mathbb{R}^m$, the algorithm needs $O\left(kmn^2\right)$ operations. We denote the application of the algorithm to $\mathcal{V}$ and $\mathcal{U}$ by*

$$[s_1, s_2, \ldots, s_m] = \textsc{DualSet}(\mathcal{V}, \mathcal{U}, k).$$

We refer the reader to [10] for the full description of the algorithm. Lemma 3.4 implies that one can sample from two different set of vectors $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ and $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$, and control *simultaneously* the smallest singular value of the matrix formed from the sampled vectors from the first set, and the largest singular value of the matrix formed from the sampled vectors from the second set. We now present our result.

**Input:** $\mathbf{X} \in \mathbb{R}^{n\times m}$, sampling parameter $n \le k \le m$.
**Output:** Set $\mathcal{S} \subseteq [m]$ of cardinality at most $k$.

1: Compute the matrix $\mathbf{V} \in \mathbb{R}^{m\times n}$ of the top $n$ right singular vectors of $\mathbf{X}$.
2: Let $\mathcal{V} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ (see Section 2 for the definition of $\mathbf{y}_i$'s).
3: Let $\mathcal{U} = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ contain the standard basis vectors.
4: Run $[s_1, s_2, \ldots, s_m] = \textsc{DualSet}(\mathcal{V},\mathcal{U},k)$ (Lemma 3.4).
5: **return** $\mathcal{S} = \{i \,:\, s_i \neq 0\}$

**Algorithm 3**: A deterministic greedy selection algorithm for subset selection (Theorem 3.5.)

**Theorem 3.5.** *Fix $\mathbf{X} \in \mathbb{R}^{n\times m}$ ($m > n$, $\mathrm{rank}(\mathbf{X}) = n$) and sampling parameter $m \ge k > n$. Algorithm 3 needs $O\left(kmn^2\right)$ operations and deterministically constructs a set $\mathcal{S} \subseteq [m]$ of cardinality at most $k$ such that for both $\xi = 2, \mathrm{F}$:*

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^2 \le \left(1 + \sqrt{\frac{m}{k}}\right)^2 \left(1 - \sqrt{\frac{n}{k}}\right)^{-2} \|\mathbf{X}^{\dagger}\|_{\xi}^2.$$

*Proof.* We first state the algorithm; then, we prove the approximation bound. Finally, we bound the number of operations.

**Algorithm.** Algorithm 3 proceeds as follows. First, it computes the SVD of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ (see also Section 2 for useful notation).. The second step is to apply the algorithm of Lemma 3.4 on $\mathcal{V} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ and $\mathcal{U} = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$, the standard basis, to compute the weights $s_1, \ldots, s_m$. The algorithm then returns $\mathcal{S} = \{i \,:\, s_i \neq 0\}$.

**Approximation Bound.** Lemma 3.4 guarantees that

$$\sigma_1\left(\sum_{i=1}^{m} s_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}}\right) \le \left(1 + \sqrt{\frac{m}{k}}\right)^2.$$

However, $\sum_{i=1}^{m} s_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} = \mathrm{diag}(s_1, \ldots, s_m) \in \mathbb{R}^{m\times m}$, a diagonal matrix containing the weights $s_i$'s in its main diagonal; so, $\max_i s_i \le \left(1 + \sqrt{\frac{m}{k}}\right)^2$. Lemma 3.4 also guarantees that

$$\sigma_n\left(\sum_{i=1}^{m} s_i \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}}\right) \ge \left(1 - \sqrt{\frac{n}{k}}\right)^2.$$

Assume that $\mathcal{S} = \{i_1, \ldots, i_{\tilde{k}}\}$ where $\tilde{k} \le k$ and $i_1 < i_2 < \cdots < i_{\tilde{k}}$, and let $\mathbf{D} = \mathrm{diag}(\sqrt{s_{i_1}}, \ldots, \sqrt{s_{i_{\tilde{k}}}})$. It is easy to verify that $\sum_{i=1}^{m} s_i \mathbf{y}_i \mathbf{y}_i^{\mathrm{T}} = \mathbf{Y}_{\mathcal{S}} \mathbf{D}^2 \mathbf{Y}_{\mathcal{S}}^{\mathrm{T}}$; so, $\mathbf{Y}_{\mathcal{S}}$ is full rank and $\|(\mathbf{Y}_{\mathcal{S}}\mathbf{D})^{\dagger}\|_2^2 \le \left(1 - \sqrt{\frac{n}{k}}\right)^{-2}$. The bound $\max_i s_i \le \left(1 + \sqrt{\frac{m}{k}}\right)^2$ earlier implies that $\|\mathbf{D}\|_2^2 \le \left(1 + \sqrt{\frac{m}{k}}\right)^2$. Now, observe that,

$$\begin{aligned}
\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^2 \overset{(a)}{=} \|\left(\mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}}\right)^{\dagger}\|_{\xi}^2 \overset{(b)}{=} \|\mathbf{Y}_{\mathcal{S}}^{\dagger}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}\|_{\xi}^2 \overset{(c)}{\le} \|\mathbf{Y}_{\mathcal{S}}^{\dagger}\|_2^2 \cdot \|\mathbf{X}^{\dagger}\|_{\xi}^2 \quad &\overset{(d)}{=} \quad \|\left(\mathbf{Y}_{\mathcal{S}}\mathbf{D}\mathbf{D}^{-1}\right)^{\dagger}\|_2^2 \cdot \|\mathbf{X}^{\dagger}\|_{\xi}^2 \\
&\overset{(e)}{\le} \quad \|\mathbf{D}\|_2^2 \cdot \|\left(\mathbf{Y}_{\mathcal{S}}\mathbf{D}\right)^{\dagger}\|_2^2 \cdot \|\mathbf{X}^{\dagger}\|_{\xi}^2 \\
&\overset{(f)}{\le} \quad \left(1 + \sqrt{\frac{m}{k}}\right)^2 \cdot \left(1 - \sqrt{\frac{n}{k}}\right)^{-2} \cdot \|\mathbf{X}^{\dagger}\|_{\xi}^2
\end{aligned}$$

**Input:** $\mathbf{X} \in \mathbb{R}^{n\times m}$, sampling parameter $n \leq k \leq m$.
**Output:** Set $\mathcal{S} \subseteq [m]$ of cardinality at most $k$.

1: Compute the matrix $\mathbf{V} \in \mathbb{R}^{m\times n}$ of the top $n$ right singular vectors of $\mathbf{X}$.
2: For $i = 1, 2, \ldots, m$ let (see Section 2 for the definition of $\mathbf{y}_i$'s),

$$\tau_i = \max\{\|\mathbf{y}_i\|_2^2, \frac{n}{m}\}.$$

3: **for** $t = 1, 2, \ldots, k$ **do**
4: Pick $i_t$; where $i_t = i$ with probability $\tau_i$.
5: **end for**
6: **return** $\mathcal{S} = \{i_1, i_2, \ldots, i_t\}$

**Algorithm 4**: A randomized algorithm for subset selection (Theorem 3.7.)

$(a)$ follows by replacing $\mathbf{X}$ with its SVD. $(b)$ follows by using Lemma 2.1 and the fact that all three matrices involved are full rank. $(c)$ follows by standard properties of matrix norms, and using the definition of the pseudoinverse of $\mathbf{X}$ and $\mathbf{\Sigma}$. $(d)$ follows by introducing the identity matrix $\mathbf{I}_{\tilde{k}} = \mathbf{D}\mathbf{D}^{-1}$. $(e)$ follows by using Lemma 2.2. Finally, $(f)$ follows from the bounds we just proved for the terms $\|\mathbf{D}\|_2^2$, and $\| \left(\mathbf{Y}_{\mathcal{S}}\mathbf{D}\right)^{\dagger} \|_2^2$.

**Operation count.** The algorithm first computes an SVD of $\mathbf{X}$, which costs $O\left(mn^2\right)$. The second step is to run the algorithm of Lemma 3.4 on the right singular vectors of $\mathbf{X}$ and the standard basis, which costs $O\left(kmn^2\right)$. So the total cost is $O\left(kmn^2\right)$. ■

### 3.3 Randomized Selection

The main idea in the algorithm of this section is to non-uniformly sample columns from $\mathbf{X}$. The analysis is based on a matrix concentration bound from [47]. More specifically, we use Theorem 3.1 from [47] (the constants are from Corollary 4 in [55]).

**Lemma 3.6** (Theorem 3.1 in [47])**.** *Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector, which is uniformly bounded almost everywhere: $\|\mathbf{x}\|_2 \leq M$. Assume, for normalization, that $\|\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]\|_2 \leq 1$. Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$ be $k$ independent copies of $\mathbf{x}$ sampled with replacement. Then, for every $\epsilon \in (0, 1)$, and with probability at least $1 - 2 \cdot n \cdot e^{-\epsilon^2 k/4M^2}$: $\|\frac{1}{k}\sum_{i=1}^{k} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]\|_2 \leq \epsilon$.*

Our algorithm is based on non-uniform sampling with replacement. The sampling probabilities are related to the so-called leverage scores of the columns of $\mathbf{X}$ [12, 25], but in our algorithm we make sure that no column has a sampling probability that is too small. One needs cubic time to compute these probabilities using SVD or QR; our algorithm computes the probabilities that way. However, one can approximate these probabilities in sub-cubic time using recent results from [24]. It might be the case that these results be used to improve the running time of our algorithm, at the cost of some small increase in the approximation bound. However, we leave this issue for future research.

**Theorem 3.7.** *Fix $\mathbf{X} \in \mathbb{R}^{n\times m}$ ($m > n$, $\mathrm{rank}(\mathbf{X}) = n$). Choose a probability parameter $\delta$ ($0 < \delta < 1$). Now, choose an sampling parameter $m \geq k \geq \min(\lceil 32n\ln(2n/\delta)\rceil, m)$. Algorithm 4 needs $O\left(mn^2 + k\log k\right)$ operations and randomly constructs a set $\mathcal{S} \subseteq [m]$ with cardinality at most $k$, such that,*

*for both $\xi = 2, \mathrm{F}$, and with probability at least $1-\delta$,*

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^{2} \leq 4 \cdot m \cdot \|\mathbf{X}^{\dagger}\|_{\xi}^{2}.$$

*Moreover, if $\mathbf{X}$ contains orthonormal rows the operation count is $O(mn + k\log k)$.*

*Proof.* We first state the algorithm; then, we prove the approximation bound. Finally, we bound the number of operations.

**Algorithm.** Algorithm 4 proceeds as follows. First, it computes the SVD of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ (see also Section 2 for useful notation). Let,

$$\tau_i = \max\{\|\mathbf{y}_i\|_2^2, \frac{n}{m}\}$$

for $i = 1, \ldots, m$. The set $\mathcal{S}$ is formed by non-uniformly, and independently, sampling $k$ numbers from $1, \ldots, m$ with replacement. In each trial, $j$ is sampled with probability

$$p_j = \tau_j / \sum_{i=1}^{m} \tau_i.$$

**Approximation bound.** The first part of the proof is a technical manipulation to enable us to use Lemma 3.6. First, let $c_1, \ldots, c_k$ be the indices sampled in trials $1, \ldots, k$. That is $\mathcal{S} = \{c_1, \ldots, c_k\}$. For $i = 1, \ldots, k$ define the random vector $\mathbf{x}_i = \mathbf{y}_{c_i}/\sqrt{p_{c_i}}$. Now, for $i = 1, \ldots, k$ define $s_i = \frac{1}{kp_i} \cdot \#\{j : c_j = i\}$. Notice that $\mathcal{S} = \{i : s_i \neq 0\}$. Assume that $\mathcal{S} = \{i_1, \ldots, i_{\tilde{k}}\}$ where $\tilde{k} \leq k$ and $i_1 < i_2 < \cdots < i_{\tilde{k}}$, and let $\mathbf{D} = \mathrm{diag}(\sqrt{s_{i_1}}, \ldots, \sqrt{s_{i_{\tilde{k}}}})$. With these definitions we observe that $\frac{1}{k}\sum_{i=1}^{k} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \mathbf{Y}_{\mathcal{S}}\mathbf{D}^2\mathbf{Y}_{\mathcal{S}}^{\mathrm{T}}$.

Notice that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are i.i.d. Let $\mathbf{x}$ denote a random vector from the same distribution of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. To use Lemma 3.6 we need to compute $\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]$ and to bound $\|\mathbf{x}\|_2^2$:

$$\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] = \sum_{i=1}^{m} p_i \cdot \frac{1}{\sqrt{p_i}}\mathbf{y}_i \cdot \frac{1}{\sqrt{p_i}}\mathbf{y}_i^{\mathrm{T}} = \sum_{i=1}^{m} \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}} = \mathbf{I}_n\,;$$

$$\|\mathbf{x}\|_2^2 \overset{(a)}{\leq} \max_{j\in[m]} \frac{\|\mathbf{y}_j\|_2^2}{p_j} \overset{(b)}{=} \max_{j\in[m]} \frac{\sum_{i=1}^{m}\tau_i}{\tau_j}\|\mathbf{y}_j\|_2^2 \overset{(c)}{\leq} \sum_{i=1}^{m}\tau_i \overset{(d)}{=} \sum_{i=1}^{m}\max\{\|\mathbf{y}_i\|_2^2, \frac{n}{m}\} \overset{(e)}{\leq} n + \sum_{i=1}^{m}\|\mathbf{y}_i\|_2^2 \overset{(f)}{=} 2n$$

$(a)$ follows by replacing the values taken by the vector $\mathbf{x}$. $(b)$ follows by replacing the value for the probabilities $p_i$'s. $(c)$ follows by the fact that $\tau_j \geq \|\mathbf{y}_j\|_2^2$, for all $j = 1, ..., m$. $(d)$ follows by replacing the value for the parameters $\tau_i$'s. $(e)$ follows by simple algebra.

We are now ready to apply Lemma 3.6 for the random vector $\mathbf{y}$ described above. An immediate application of this Lemma ($M = \sqrt{2n}$, $\epsilon = 1/2$) and our bound on $k$ give that with probability at least $1-\delta$, $\|\mathbf{Y}_{\mathcal{S}}\mathbf{D}^2\mathbf{Y}_{\mathcal{S}}^{\mathrm{T}} - \mathbf{I}_n\|_2 \leq \frac{1}{2}$. (Recall that $\frac{1}{k}\sum_{i=1}^{k}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} = \mathbf{Y}_{\mathcal{S}}\mathbf{D}^2\mathbf{Y}_{\mathcal{S}}^{\mathrm{T}}$). Standard matrix perturbation theory results [29] imply that for $i = 1, ..., n$ $|\sigma_i^2(\mathbf{Y}_{\mathcal{S}}\mathbf{D}) - 1| \leq \|\mathbf{Y}_{\mathcal{S}}\mathbf{D}^2\mathbf{Y}_{\mathcal{S}}^{\mathrm{T}} - \mathbf{I}_n\|_2$, so, $i = n$ gives, $\|(\mathbf{Y}_{\mathcal{S}}\mathbf{D})^{\dagger}\|_2^2 \leq 2$. We bound $\|\mathbf{D}\|_2^2$ as follows

$$\begin{aligned}
\|\mathbf{D}\|_2^2 \overset{(a)}{=} \max_{j\in[m]} \mathbf{D}_{jj}^2 \overset{(b)}{\leq} k \max_{j\in[m]}\left(\frac{1}{kp_j}\right) \overset{(c)}{\leq} \max_{j\in[m]}\left(\frac{\sum_{i=1}^{m}\tau_i}{\tau_j}\right) &\overset{(d)}{=} \left(\sum_{i=1}^{m}\tau_i\right)\cdot\max_{j\in[m]}\left(\frac{1}{\max\{\|\mathbf{y}_j\|_2^2, n/m\}}\right) \\
&\overset{(e)}{\leq} 2\cdot n\cdot\max_{j\in[m]}\left(\frac{1}{\max\{\|\mathbf{y}_j\|_2^2, n/m\}}\right) \\
&\overset{(f)}{\leq} 2\cdot n\cdot\frac{m}{n} = 2\cdot m
\end{aligned}$$

$(a)$ follows because $\mathbf{D}^2$ is a diagonal matrix. $(b)$ follows because each entry in $\mathbf{D}^2$ might contain the term $1/kp_j$, at most $k$ times. $(c)$ follows by replacing the values for the probabilities. $(d)$ follows by replacing the values of the parameters $\tau_j$'s. $(e)$ follows by the fact that $\sum_{i=1}^{m} \tau_i \leq 2n$, which we proved in Eqn. $(f)$ in the previous calculations. $(f)$ follows by simple algebra.

To conclude the proof, notice that at the end of Lemma 3.5, we implicitly proved that $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\xi}^2 \leq \|\mathbf{D}\|_2^2 \cdot \|(\mathbf{Y}_{\mathcal{S}}\mathbf{D})^{\dagger}\|_2^2 \cdot \|\mathbf{X}^{\dagger}\|_{\xi}^2$. Replace the bounds for $\|(\mathbf{Y}_{\mathcal{S}}\mathbf{D})^{\dagger}\|_2^2$ and $\|\mathbf{D}\|_2^2$ in this bound to wrap up.

**Operation count.** First, we need to compute an SVD of $\mathbf{X}$, which costs $O\left(mn^2\right)$. The probabilities can be calculated in $O(mn)$ and the sampling procedure can be implemented in $O(m + k\log k)$. In total, the cost is $O\left(mn^2 + k\log k\right)$. If $\mathbf{X}$ contains orthonormal rows, $O(mn + k\log k)$ operations suffice. ■

### 3.4 Volume based bounds and algorithms

We now consider bounds and algorithms which construct the set $\mathcal{S}$ by looking at the volume of the parallelepiped spanned by the rows of $\mathbf{X}_{\mathcal{S}}$, which is exactly the determinant of $\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}$.

#### Subset Selection and Determinants

We start with Lemma 3.8, which establishes the connection between determinants and subset selection.

**Lemma 3.8.** *Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ $(m \geq n)$ be a full rank matrix, and let $\mathcal{S} \subseteq [m]$ be any subset of cardinality $k$ $(n \leq k \leq m)$ such that $\mathbf{X}_{\mathcal{S}}$ is full rank. For $i = 1, \ldots, n$, let $\mathbf{Y}_i \in \mathbb{R}^{(n-1)\times m}$ denote the matrix obtained after removing the $i$th row of $\mathbf{X}$. Then,*

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 = \frac{\sum_{i=1}^{n} \det\left((\mathbf{Y}_i)_{\mathcal{S}} (\mathbf{Y}_i)_{\mathcal{S}}^{\mathrm{T}}\right)}{\det\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)}.$$

*Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]$ be the column representation of $\mathbf{X}$. If $\mathcal{S}$ has cardinality exactly $n$, then*

$$\|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2 \leq \|\mathbf{X}^{\dagger}\|_2^2 \cdot \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} \det\left(\mathbf{X}_{\mathcal{S}}(i \to \mathbf{x}_j)\right)^2}{\det\left(\mathbf{X}_{\mathcal{S}}\right)^2}.$$

*Recall that, $\mathbf{X}_{\mathcal{S}}\left(i \to \mathbf{x}_j\right)$ is the matrix by replacing the $i$-th column of $\mathbf{X}_{\mathcal{S}}$ with $\mathbf{x}_j$, the $j$-th column of $\mathbf{X}$. If $\mathbf{X}$ has orthonormal rows then the last inequality is an equality.*

*Proof.* We first prove the equality in the Lemma ($\mathcal{S}$ has cardinality k unless otherwise stated),

$$\begin{aligned}
\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \overset{(a)}{=} \mathrm{Tr}\left(\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)^{-1}\right) \overset{(b)}{=} \mathrm{Tr}\left(\det\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)^{-1} \mathrm{Adj}\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)\right) &\overset{(c)}{=} \det\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)^{-1} \mathrm{Tr}\left(\mathrm{Adj}\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)\right) \\
&\overset{(d)}{=} \det\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \left(\mathrm{Adj}\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)\right)_{ii} \\
&\overset{(e)}{=} \det\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{n} \det\left((\mathbf{Y}_i)_{\mathcal{S}} (\mathbf{Y}_i)_{\mathcal{S}}^{\mathrm{T}}\right)
\end{aligned}$$

$(a)$ follows by a property which connects the Frobenius norm of the pseudoinverse with the trace operator. $(b)$ follows by the well known formula for the inverse of a matrix using the adjugate matrix. $(c)$ follows by the linearity of the trace operator. $(d)$ follows by the definition of the trace operator. Finally, $(e)$ follows by the definition of the adjugate matrix and the observation that the $i$th diagonal element of $\mathrm{Adj}\left(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}\right)$

equals the determinant of an $(n-1)\times(n-1)$ matrix which is exactly $\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}$ after removing its $i$th row and $i$th column. This matrix is exactly $(\mathbf{Y}_i)_{\mathcal{S}}(\mathbf{Y}_i)_{\mathcal{S}}^{\mathrm{T}}$.

We now prove the second inequality ($\mathcal{S}$ has now fixed cardinality $k$). Let $\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be an SVD of $\mathbf{X}$ (see also Section 2 for useful notation). Define $\mathbf{x}_j=\mathbf{U}\mathbf{\Sigma}\mathbf{y}_j$. Then,

$$\begin{aligned}
\|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2 \overset{(a)}{=} \|\mathbf{Y}_{\mathcal{S}}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}\|_{\mathrm{F}}^2 \overset{(b)}{\leq} \|\mathbf{\Sigma}^{-1}\|_2^2\cdot\|\mathbf{Y}_{\mathcal{S}}^{-1}\mathbf{Y}\|_{\mathrm{F}}^2 &\overset{(c)}{=} \|\mathbf{\Sigma}^{-1}\|_2^2\cdot\sum_{j=1}^{m}\|\mathbf{Y}_{\mathcal{S}}^{-1}\mathbf{y}_j\|_2^2 \\
&\overset{(d)}{=} \|\mathbf{\Sigma}^{-1}\|_2^2\cdot\sum_{j=1}^{m}\|\mathbf{Y}_{\mathcal{S}}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{y}_j\|_2^2 \\
&\overset{(e)}{=} \|\mathbf{\Sigma}^{-1}\|_2^2\cdot\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\det\left((\mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}})\,(i\rightarrow\mathbf{U}\mathbf{\Sigma}\mathbf{y}_j)\right)^2}{\det(\mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}})^2} \\
&\overset{(f)}{=} \|\mathbf{X}^{\dagger}\|_2^2\cdot\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\det(\mathbf{X}_{\mathcal{S}}(i\rightarrow\mathbf{x}_j))^2}{\det(\mathbf{X}_{\mathcal{S}})^2}\,.
\end{aligned}$$

$(a)$ follows by replacing the SVD of $\mathbf{X}_{\mathcal{S}}^{-1}$. Notice that $\mathbf{X}_{\mathcal{S}}=\mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}}$ and $\mathbf{X}_{\mathcal{S}}^{-1}=\mathbf{Y}_{\mathcal{S}}^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}$. The latter equality holds because $\mathbf{Y}_{\mathcal{S}}^{-1}$ is a square full rank matrix, which is immediate by the assumption that $\mathrm{rank}(\mathbf{X}_{\mathcal{S}})=n$. $(b)$ follows by first using a property of matrix norms, and, then, dropping the square orthonormal matrix $\mathbf{U}^{\mathrm{T}}$ and inserting the matrix $\mathbf{Y}$, which has orthonormal rows, to the Frobenius norm term. $(c)$ follows by the definition of the Frobenius norm. $(d)$ follows by inserting the identity matrix $\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}=\mathbf{I}_n$. $(e)$ follows by applying Cramer's rule to the linear system $\mathbf{A}\mathbf{x}=\mathbf{b}$, with $\mathbf{A}=\mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}}$ and $\mathbf{b}=\mathbf{U}\mathbf{\Sigma}\mathbf{y}_j$. Finally, $(f)$ follows by replacing the appropriate values for $\mathbf{X}_{\mathcal{S}}$, $\mathbf{x}_j$, and $\|\mathbf{\Sigma}^{-1}\|_2^2$.

Notice that if $\mathbf{X}$ has orthonormal rows then $\mathbf{\Sigma}$ is the identity matrix, and $(b)$ becomes an equality. ■

**Random Subsets Chosen via Volume Sampling**

Lemma 3.8 connects determinants and the term $\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2$, for *any* set $\mathcal{S}$ for which $\mathbf{X}_{\mathcal{S}}$ has full rank. In the related work part of the introduction, we also stated various results for the specific set $\hat{\mathcal{S}}\subseteq[m]$ of cardinality $k\geq n$ that maximizes $\det(\mathbf{X}_{\mathcal{T}}\mathbf{X}_{\mathcal{T}}^{\mathrm{T}})$ over all possible $\mathcal{T}$'s of cardinality $k$. Unfortunately, finding the maximum volume (determinant) subset is not only NP-hard [44, 19], but also exponentially hard to approximate [40, 18], so, these results do not yield an efficient algorithm. We solve this issue using randomization.

**Lemma 3.9.** *Let $\mathbf{X}\in\mathbb{R}^{n\times m}$ $(m\geq n)$ be a full rank matrix and let $m\geq k\geq n$. Suppose that $\mathcal{S}\sim\mathrm{VolSamp}(\mathbf{X},k)$. Then,*

$$\mathbb{E}\left[\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2\right]\leq\frac{m-n+1}{k-n+1}\cdot\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2\,. \tag{2}$$

*(If for every set $\mathcal{S}\in C([m],n)$ the matrix $\mathbf{X}_{\mathcal{S}}$ is full rank then this bound becomes an equality). Also, for $i=1,\ldots,n$,*

$$\mathbb{E}\left[\sigma_i^{-2}(\mathbf{X}_{\mathcal{S}})\right]\leq\left(1+\frac{n(m-k)}{k-n+1}\right)\cdot\sigma_i^{-2}(\mathbf{X})\,.$$

*In particular,*

$$\mathbb{E}\left[\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2\right]\leq\left(1+\frac{n(m-k)}{k-n+1}\right)\cdot\|\mathbf{X}^{\dagger}\|_2^2\,.$$

*Proof.* For $i = 1, \dots, n$, let $\mathbf{Y}_i \in \mathbb{R}^{(n-1)\times m}$ denote the matrix obtained after removing the $i$th row of $\mathbf{X}$. Using the definition of expectation and the equality of Lemma 3.8,

$$\mathbb{E}\left[\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2\right] = \frac{\sum_{\mathcal{S}\in C([m],k)} \det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}})\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2}{\sum_{\mathcal{S}\in C([m],k)} \det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}})} \overset{(*)}{=} \frac{\sum_{\mathcal{S}\in C([m],k),\mathrm{rank}(\mathbf{X}_{\mathcal{S}})=n} \det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}})\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2}{\sum_{\mathcal{S}\in C([m],k)} \det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}})}$$
$$= \frac{\sum_{\mathcal{S}\in C([m],k),\mathrm{rank}(\mathbf{X}_{\mathcal{S}})=n} \sum_{i=1}^{n} \det((\mathbf{Y}_i)_{\mathcal{S}}(\mathbf{Y}_i)_{\mathcal{S}}^{\mathrm{T}})}{\sum_{\mathcal{S}\in C([m],k)} \det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}})}.$$

In $(*)$, if $\mathrm{rank}(\mathbf{X}_{\mathcal{S}}) \neq n$, then, $\det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}) = 0$, so it can be ignored in the sum. We will now analyze the numerator and the denominator of the last relation separately. We start with the denominator. We have

$$\sum_{\mathcal{S}\in C([m],k)} \det(\mathbf{X}_{\mathcal{S}}\mathbf{X}_{\mathcal{S}}^{\mathrm{T}}) \overset{(a)}{=} \sum_{\mathcal{S}\in C([m],k)} \sum_{\mathcal{T}\in C(\mathcal{S},n)} \det(\mathbf{X}_{\mathcal{T}}\mathbf{X}_{\mathcal{T}}^{\mathrm{T}}) \overset{(b)}{=} \binom{m-n}{k-n} \sum_{\mathcal{T}\in C([m],n)} \det(\mathbf{X}_{\mathcal{T}}\mathbf{X}_{\mathcal{T}}^{\mathrm{T}}) \overset{(c)}{=} \binom{m-n}{k-n} \prod_{i=1}^{n} \sigma_i^2,$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ are the singular values of $\mathbf{X}$. $(a)$ follows by applying the Cauchy-Binet formula. $(b)$ follows from observing that each set in $\binom{[m]}{n}$ is repeated exactly $\binom{m-n}{k-n}$ times in the sum. $(c)$ follows by applying the Cauchy-Binet formula again and the fact that for symmetric positive-definite matrices the determinant is equal to the product of the eigenvalues.

As for the numerator, we have

$$\begin{aligned}
\sum_{\mathcal{S}\in C([m],k),\mathrm{rank}(\mathbf{X}_{\mathcal{S}})=n} \sum_{i=1}^{n} \det((\mathbf{Y}_i)_{\mathcal{S}}(\mathbf{Y}_i)_{\mathcal{S}}^{\mathrm{T}}) &\overset{(a)}{\leq} \sum_{\mathcal{S}\in C([m],k)} \sum_{i=1}^{n} \det((\mathbf{Y}_i)_{\mathcal{S}}(\mathbf{Y}_i)_{\mathcal{S}}^{\mathrm{T}}) \\
&= \sum_{i=1}^{n} \sum_{\mathcal{S}\in C([m],k)} \det((\mathbf{Y}_i)_{\mathcal{S}}(\mathbf{Y}_i)_{\mathcal{S}}^{\mathrm{T}}) \\
&\overset{(b)}{=} \sum_{i=1}^{n} \sum_{\mathcal{S}\in C([m],k)} \sum_{\mathcal{T}\in C(\mathcal{S},n-1)} \det((\mathbf{Y}_i)_{\mathcal{T}}(\mathbf{Y}_i)_{\mathcal{T}}^{\mathrm{T}}) \\
&\overset{(c)}{=} \sum_{i=1}^{n} \binom{m-n+1}{k-n+1} \sum_{\mathcal{T}\in C([m],n-1)} \det((\mathbf{Y}_i)_{\mathcal{T}}(\mathbf{Y}_i)_{\mathcal{T}}^{\mathrm{T}}) \\
&\overset{(d)}{=} \binom{m-n+1}{k-n+1} \sum_{i=1}^{n} \det(\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}) \\
&\overset{(e)}{=} \binom{m-n+1}{k-n+1} \sum_{i=1}^{n} \prod_{j\neq i} \sigma_i^2.
\end{aligned}$$

$(a)$ follows because we are adding only positive terms in the sum. $(b)$ follows by applying the Cauchy-Binet formula. $(c)$ follows from observing that each set in $C([m], n-1)$ is repeated exactly $\binom{m-n+1}{k-n+1}$ times in the sum. $(d)$ follows by applying the Cauchy-Binet formula again. Finally, in $(e)$, the matrices $\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}$ are equal to the matrix obtained by deleting the $i$-th column and the $i$-th row of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, so according to Lemma 2.4, $\sum_{i=1}^{n} \det(\mathbf{Y}_i\mathbf{Y}_i^{\mathrm{T}}) = \sum_{i=1}^{n} \prod_{j\neq i} \sigma_i^2$.

We now conclude the first part of the proof as follows,

$$\mathbb{E}\left[\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2\right] \leq \frac{\binom{m-n+1}{k-n+1} \sum_{i=1}^{n} \prod_{j\neq i} \sigma_i^2}{\binom{m-n}{k-n} \prod_{i=1}^{n} \sigma_i^2} = \frac{\binom{m-n+1}{k-n+1} \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2}{\binom{m-n}{k-n}} = \frac{m-n+1}{k-n+1} \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2.$$

(If $\mathbf{X}_\mathcal{S}$ is full rank for every $\mathcal{S} \in C([m], n)$ then $(a)$ in the previous calculations is an equality.)

We now prove the bounds for the singular values of $\mathbf{X}_\mathcal{S}$. Let $\mathcal{T}$ be any subset of $[m]$ of cardinality $k$ such that $\mathbf{X}_\mathcal{T}$ has full rank, and let $\bar{\mathcal{T}} = [m] - \mathcal{T}$. Notice that $\bar{\mathcal{T}}$ has cardinality $m-k$. Let,

$$\mathbf{W} = \begin{pmatrix} \mathbf{I}_k & \mathbf{X}_\mathcal{T}^\dagger \mathbf{X}_{\bar{\mathcal{T}}} \\ \mathbf{0}_{(m-k)\times k} & \mathbf{I}_{m-k} \end{pmatrix} \in \mathbb{R}^{m\times m}.$$

(Note that $\mathbf{X}_\mathcal{T}^\dagger \mathbf{X}_{\bar{\mathcal{T}}} \in \mathbb{R}^{k\times(m-k)}$). Since $\mathbf{X}_\mathcal{T}$ has full rank, we have

$$\begin{pmatrix} \mathbf{X}_\mathcal{T} & \mathbf{0}_{n\times(m-k)} \end{pmatrix} \mathbf{W} = \begin{pmatrix} \mathbf{X}_\mathcal{T} & \mathbf{X}_{\bar{\mathcal{T}}} \end{pmatrix} = \mathbf{X}\mathbf{\Pi},$$

where $\mathbf{\Pi} \in \mathbb{R}^{m\times m}$ is an appropriate permutation matrix. Clearly $\mathbf{W}$ is non-singular (it is a triangular matrix with a non-zero diagonal), so for $i = 1, \ldots, n$, a simple matrix perturbation argument implies that,

$$\sigma_i^{-2}(\mathbf{X}_\mathcal{T}) = \sigma_i^{-2}\left(\begin{pmatrix} \mathbf{X}_\mathcal{T} & \mathbf{0}_{n\times(m-k)} \end{pmatrix}\right) = \sigma_i^{-2}\left(\mathbf{X}\mathbf{\Pi}\mathbf{W}^{-1}\right) \leq \|\mathbf{W}\|_2^2 \cdot \sigma_i^{-2}(\mathbf{X}).$$

To bound $\|\mathbf{W}\|_2^2$ we observe that,

$$\|\mathbf{W}\|_2^2 \leq 1 + \|\mathbf{X}_\mathcal{T}^\dagger \mathbf{X}_{\bar{\mathcal{T}}}\|_2^2 \leq 1 + \|\mathbf{X}_\mathcal{T}^\dagger \mathbf{X}_{\bar{\mathcal{T}}}\|_\mathrm{F}^2.$$

Now, if $\mathcal{S} \sim \mathrm{VolSamp}(\mathbf{X}, k)$ then only $\mathcal{S}$'s for which $\mathbf{X}_\mathcal{S}$ is full rank have positive probability of being sampled. This implies that for $i = 1, \ldots, n$,

$$\mathbb{E}\left[\sigma_i^{-2}(\mathbf{X}_\mathcal{S})\right] \leq \mathbb{E}\left[\left(1 + \|\mathbf{X}_\mathcal{S}^\dagger \mathbf{X}_{\bar{\mathcal{S}}}\|_\mathrm{F}^2\right)\right] \cdot \sigma_i^{-2}(\mathbf{X}).$$

We now bound $\mathbb{E}\left[\left(1 + \|\mathbf{X}_\mathcal{S}^\dagger \mathbf{X}_{\bar{\mathcal{S}}}\|_\mathrm{F}^2\right)\right]$. Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathrm{T}$ be a SVD of $\mathbf{X}$ and let us denote $\mathbf{Z} = \mathbf{V}^\mathrm{T}$. Then it is easy to verify that $\mathbf{X}_\mathcal{S}^\dagger \mathbf{X}_{\bar{\mathcal{S}}} = \mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_{\bar{\mathcal{S}}}$. To bound the expected value of $\|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_{\bar{\mathcal{S}}}\|_\mathrm{F}^2$ we observe that,

$$\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}\mathbf{\Pi} = \begin{pmatrix} \mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_\mathcal{S} & \mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_{\bar{\mathcal{S}}} \end{pmatrix}.$$

This implies that $\|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}\mathbf{\Pi}\|_\mathrm{F}^2 = \|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_\mathcal{S}\|_\mathrm{F}^2 + \|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_{\bar{\mathcal{S}}}\|_\mathrm{F}^2$. $\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_\mathcal{S}$ is a projection; so, $\|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_\mathcal{S}\|_\mathrm{F}^2 = n$. We now have $\|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}\mathbf{\Pi}\|_\mathrm{F}^2 = n + \|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_{\bar{\mathcal{S}}}\|_\mathrm{F}^2$. $\mathbf{Z}\mathbf{\Pi}$ has orthonormal rows; so, $\|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}\mathbf{\Pi}\|_\mathrm{F}^2 = \|\mathbf{Z}_\mathcal{S}^\dagger\|_\mathrm{F}^2$. So, $\|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_{\bar{\mathcal{S}}}\|_\mathrm{F}^2 = \|\mathbf{Z}_\mathcal{S}^\dagger\|_\mathrm{F}^2 - n$. Since $\mathrm{VolSamp}(\mathbf{X}, k) = \mathrm{VolSamp}(\mathbf{Z}, k)$, the Frobenius norm bound in the lemma guarantees that,

$$\mathbb{E}\left[\|\mathbf{Z}_\mathcal{S}^\dagger\|_\mathrm{F}^2\right] \leq \frac{n(m-n+1)}{k-n+1}.$$

Plugging that into the previous equality, we find that,

$$\mathbb{E}\left[\|\mathbf{Z}_\mathcal{S}^\dagger \mathbf{Z}_{\bar{\mathcal{S}}}\|_\mathrm{F}^2\right] \leq \frac{n(m-k)}{k-n+1}.$$

This immediately gives a bound on $\mathbb{E}\left[\left(1 + \|\mathbf{X}_\mathcal{S}^\dagger \mathbf{X}_{\bar{\mathcal{S}}}\|_\mathrm{F}^2\right)\right]$, which concludes the proof. ∎

We can now prove the following corollary, which was previously stated as Lemma 3.2.

**Corollary 3.10** (Restatement of Lemma 3.2)**.** *Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ $(m \geq n)$ be a full rank matrix. There exists a subset $\mathcal{S} \subset [m]$ of cardinality $m-1$ such that $\mathbf{X}_\mathcal{S}$ is full rank and*

$$\|\mathbf{X}_\mathcal{S}^\dagger\|_\mathrm{F}^2 \leq \frac{m-n+1}{m-n} \cdot \|\mathbf{X}^\dagger\|_\mathrm{F}^2.$$

**Input:** $\mathbf{X} \in \mathbb{R}^{n\times m}$, parameter $\eta > 0$.
**Output:** Set $\mathcal{S} \subseteq [m]$ of cardinality $n$.

Let $\alpha = (1+\eta)\cdot(m-n+1)\cdot\|\mathbf{X}^\dagger\|_{\mathrm{F}}^2$.
**repeat**
  Apply VolumeSample from [34] to sample a subset $\mathcal{S}$ from $\mathrm{VolSamp}(\mathbf{X}, n)$.
**until** $\|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2 \leq \alpha$
**return** $\mathcal{S}$

**Algorithm 5**: A randomized volume-based sampling algorithm for subset selection (Theorem 3.11.)

*Proof.* Let $\mathcal{T} \sim \mathrm{VolSamp}(\mathbf{X}, m-1)$. According to Lemma 3.9 we have

$$\mathbb{E}\left[\|\mathbf{X}_{\mathcal{T}}^\dagger\|_{\mathrm{F}}^2\right] \leq \frac{m-n+1}{m-n}\cdot\|\mathbf{X}^\dagger\|_{\mathrm{F}}^2 .$$

The random variable $\|\mathbf{X}_{\mathcal{T}}^\dagger\|_{\mathrm{F}}^2$ is discrete, so it must assume at least one value larger than the expectancy with non-zero probability. Let $\mathcal{S}$ be one such set, so

$$\|\mathbf{X}_{\mathcal{S}}^\dagger\|_{\mathrm{F}}^2 \leq \frac{m-n+1}{m-n}\cdot\|\mathbf{X}^\dagger\|_{\mathrm{F}}^2 .$$

$\mathbf{X}_{\mathcal{S}}$ must be full rank since $\mathcal{T}$ assumes it with some non-zero probability, but the distribution $\mathrm{VolSamp}(\mathbf{X}, m-1)$ gives a zero probability to every subset $\mathcal{R}$ for which $\mathbf{X}_{\mathcal{R}}$ is rank deficient. ∎

If there exists a subset $\mathcal{S}$ of columns of cardinality $k$ such that these columns are linearly dependent then (2) might be a strict inequality. For example, let $\mathbf{Y} = \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0}_{n\times(m-n)} \end{pmatrix}$. There is only one set of cardinality $n$ that has positive volume (i.e., the set of columns is full rank): $\mathcal{T} = [n]$. Since this is the only set with positive probability we have

$$\mathbb{E}\left[\|\mathbf{Y}_{\mathcal{S}}^\dagger\|_{\mathrm{F}}^2\right] = n = \|\mathbf{Y}^\dagger\|_{\mathrm{F}}^2 < (m-n+1)\|\mathbf{Y}_{\mathcal{S}}^\dagger\|_{\mathrm{F}}^2 = (m-n+1)n .$$

It is interesting to note that there is a discontinuity near matrices of this sort. Let $\mathbf{Z}$ be a matrix such that $\mathbf{Y}_\epsilon = \mathbf{Y} + \epsilon\mathbf{Z}$ has full rank for every subset $\mathcal{T}$ and every $\epsilon \neq 0$. Obviously such a matrix exist. It is easy to see that for $\epsilon \to 0$ we have $\|\mathbf{Y}_\epsilon^\dagger\|_{\mathrm{F}}^2 \to \|\mathbf{Y}^\dagger\|_{\mathrm{F}}^2 = n$. We find that

$$\lim_{\epsilon\to 0}\mathbb{E}\left[\|(\mathbf{Y}_\epsilon)_{\mathcal{S}}^\dagger\|_{\mathrm{F}}^2\right] = (m-n+1)n \neq n = \mathbb{E}\left[\|\mathbf{Y}^\dagger\|_{\mathrm{F}}^2\right] .$$

**Volume Sampling Subset Selection**

To use Lemma 3.9 in an algorithm one needs a method to sample a subset from $\mathrm{VolSamp}(\mathbf{X}, k)$. Computing the probabilities for all $\binom{m}{k}$ subsets and sampling according to them is not efficient; there are too many such sets. However, this is not necessary, since one can *simulate* volume sampling using polynomial number of operations using recent results from [22, 34]. More precisely, using the algorithm VolumeSample from [34] we can sample a subset $\mathcal{S}$ from $\mathrm{VolSamp}(\mathbf{X}, n)$ using $O(n^3 m)$ operations. Current determinant-based sampling algorithms [22, 34] can sample only $k = n$ columns from $\mathbf{X}$. We leave it as an open question whether one can efficiently sample from $\mathrm{VolSamp}(\mathbf{X}, k)$ for arbitrary $k \geq n$.

**Theorem 3.11.** *Fix $\mathbf{X} \in \mathbb{R}^{n\times m}$ ($m \geq n$, $\operatorname{rank}(\mathbf{X}) = n$) and sampling parameter $k = n$. Choose a parameter $\eta > 0$. Algorithm 5 constructs a set $\mathcal{S} \subseteq [m]$ of cardinality $k$ with,*

$$\|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2 \leq (1+\eta)\cdot(m-n+1)\cdot\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2; \quad \|\mathbf{X}_{\mathcal{S}}^{-1}\|_2^2 \leq (1+\eta)\cdot(m-n+1)\cdot n\cdot\|\mathbf{X}^{\dagger}\|_2^2.$$

*For every $0 < \delta < 1$, the algorithm will terminate after $O\left(mn^3\log(1/\delta)/\log(1+\eta)\right)$ operations with probability at least $1-\delta$.*

*Proof.* First of all, we only need to prove the bound for the Frobenius norm. The bounds for the spectral norm follow by the fact that for any matrix $\mathbf{B}$ of rank $n$: $\|\mathbf{B}\|_2^2 \leq \|\mathbf{B}\|_{\mathrm{F}}^2 \leq n\|\mathbf{B}\|_2^2$.

**Algorithm.** We describe Algorithm 5. We start by applying VolumeSample from [34] to sample a subset $\mathcal{S}$ from $\operatorname{VolSamp}(\mathbf{X}, n)$ using $O(n^3m)$ operations. We then compute $\|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2$ using $O(n^3)$ operations and compare it to $(1+\eta)\cdot(m-n+1)\cdot\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2$. If it is smaller than the bound we return $\mathcal{S}$, otherwise we repeat this procedure until we find a satisfactory $\mathcal{S}$.

**Approximation bound.** Notice that we repeat the experiment $t = 1, 2, ...$ times, constructing sets $\mathcal{S}_1$, $\mathcal{S}_2$.... We stop only if we find a satisfactory $\mathcal{S}$, so the bounds are satisfied if the algorithm returns.

**Failure probability.** Lemma 3.9 indicates that if $\mathcal{S}$ is sampled from $\operatorname{VolSamp}(\mathbf{X}, n)$, then, $\mathbb{E}\left[\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2\right] \leq (m-n+1)\cdot\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2$. For the first iteration $t = 1$, by Markov's inequality, we find that, with probability at most $1/(1+\eta)$: $\|\mathbf{X}_{\mathcal{S}_1}^{\dagger}\|_{\mathrm{F}}^2 > (1+\eta)(m-n+1)\cdot\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2$. Therefore, for a finite number of iterations $\ell > 1$, the probability that all $t = 1, ..., \ell$, satisfy $\|\mathbf{X}_{\mathcal{S}_t}^{\dagger}\|_{\mathrm{F}}^2 > (1+\eta)(m-n+1)\cdot\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2$ is at most $1/(1+\eta)^{\ell}$. So, for any $0 < \delta < 1$ the probability $\lceil \log(1/\delta)/\log(1+\eta) \rceil$ successive iterations to fail is below $\delta$.

**Operation count.** Each iteration of the algorithm takes $O(n^3m)$. In total, for any $0 < \delta < 1$ the operation count is $O\left(mn^3\log(1/\delta)/\log(1+\eta)\right)$ with probability $1-\delta$. ∎

# 4 Lower Bounds

## 4.1 Spectral norm subset selection

**Theorem 4.1** (Spectral Norm)**.** *For any $\alpha > 0$, $n > 0$, $m > 2$ with $m > n$, and $k$ with $n \leq k \leq m$, there is a full rank $n \times m$ matrix $\mathbf{X}$ such that, for any subset $\mathcal{S} \subseteq [m]$ of cardinality $k \geq n$ with $\operatorname{rank}(\mathbf{X}_{\mathcal{S}}) = n$,*

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2 \geq \left(\frac{m+\alpha^2}{k+\alpha^2} - 1\right)\cdot\|\mathbf{X}^{\dagger}\|_2^2$$

*Proof.* We construct the matrix $\mathbf{X}$ as follows. Let $\mathbf{A} = [\mathbf{e}_1 + \alpha\mathbf{e}_2, \mathbf{e}_1 + \alpha\mathbf{e}_3, \ldots, \mathbf{e}_1 + \alpha\mathbf{e}_{m+1}] \in \mathbb{R}^{(m+1)\times m}$, and let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be the SVD decomposition of $\mathbf{A}$. $\mathbf{X}$ is the first $n$ rows of $\mathbf{V}^{\mathrm{T}}$.

To prove the bound we use Theorem 17 and Lemma 7 from [10]. Theorem 17 shows that if $m > 2$, then, for $\mathbf{A}$ defined above, for every subset $\mathcal{S} \subseteq [m]$ of cardinality $k$, we have

$$\|\mathbf{A} - \mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^{\dagger}\mathbf{A}\|_2^2 = \frac{m+\alpha^2}{k+\alpha^2}\cdot\|\mathbf{A} - \mathbf{A}_n\|_2^2,$$

where $\mathbf{A}_n$ is the best rank $n$ approximation to $\mathbf{A}$. Lemma 7 proves that for any matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$, rank parameter $n < \operatorname{rank}(\mathbf{W})$, and sampling parameter $n \leq k \leq m$, for every subset $\mathcal{S} \subseteq [m]$ of cardinality $k$ for which $\mathbf{Z}_{\mathcal{S}}$ ($\mathbf{Z}$ is to be defined shortly) has full rank, we have

$$\|\mathbf{W} - \mathbf{W}_{\mathcal{S}}\mathbf{W}_{\mathcal{S}}^{\dagger}\mathbf{W}\|_2^2 \leq \|\mathbf{W} - \mathbf{W}_n\|_2^2 + \|(\mathbf{W} - \mathbf{W}_n)_{\mathcal{S}}\mathbf{Z}_{\mathcal{S}}^{\dagger}\|_2^2 \leq \left(1 + \|\mathbf{Z}_{\mathcal{S}}^{\dagger}\|_2^2\right)\cdot\|\mathbf{W} - \mathbf{W}_n\|_2^2.$$

Here, $\mathbf{W}_n$ is the best rank $n$ approximation to $\mathbf{W}$. $\mathbf{Z}$ is defined as follows. Let $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be the SVD of $\mathbf{W}$. $\mathbf{Z}$ is the first $n$ rows of $\mathbf{V}^{\mathrm{T}}$.

We now apply Lemma 7 from [10] on $\mathbf{A}$ and $\mathbf{X}$ and combine it with the bound from Theorem 17 of [10] mentioned above, to find that

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_2^2 \geq \frac{\|\mathbf{A} - \mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^{\dagger}\mathbf{A}\|_2^2}{\|\mathbf{A} - \mathbf{A}_n\|_2^2} - 1 = \frac{m+\alpha^2}{k+\alpha^2} - 1 = \left(\frac{m+\alpha^2}{k+\alpha^2} - 1\right)\|\mathbf{X}^{\dagger}\|_2^2.$$

■

As $\alpha \to 0$, the bound in the above theorem is $m/k - 1$. If $k = (1+\Omega(1))n$ then the upper bound of the deterministic algorithm of Theorem 3.5 asymptotically matches this lower bound. The upper bounds of the algorithms in the Theorems 3.1, 3.7, and 3.11 and Corollary 3.3 are - asymptotically - slightly worse. However, if $k = (1+o(1))n$ there is a gap between the lower bound and the best upper bound.

### 4.2 Frobenius norm subset selection

**Theorem 4.2** (Frobenius Norm)**.** *For any $\alpha > 0$, $n$, $m$ with $m > n$, $\mod(m,n) = 0$, and $m/n > 2$, and $k$ with $n \leq k \leq m$, there is a full rank $n \times m$ matrix $\mathbf{X}$ such that, for any subset $\mathcal{S} \subseteq [m]$ of cardinality $k \geq n$ with $\mathrm{rank}(\mathbf{X}_{\mathcal{S}}) = n$, we have*

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \geq \left(\frac{m-k}{k+\alpha^2} + 1 - \frac{k}{n}\right) \cdot \|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2$$

*Proof.* We construct the matrix $\mathbf{X}$ as follows. Consider a block diagonal matrix $\mathbf{B} \in \mathbb{R}^{d\times m}$: a matrix $\mathbf{A} \in \mathbb{R}^{d/n \times m/n}$ of the form that appear in the proof of Theorem 4.1 is repeated $n$ times on $\mathbf{B}$'s main diagonal. Let $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ is the SVD of $\mathbf{B}$. $\mathbf{X}$ is the first $n$ rows of $\mathbf{V}^{\mathrm{T}}$.

To prove the bound we use Theorem 19 and Lemma 7 from [10]. Theorem 19 in [10] indicates that for the above $\mathbf{B}$, any $n \leq \mathrm{rank}(\mathbf{B})$, and $k \geq n$, any subset $\mathcal{S}$ of $k$ columns of $\mathbf{B}$ satisfies,

$$\|\mathbf{B} - \mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^{\dagger}\mathbf{B}\|_{\mathrm{F}}^2 = \frac{m-k}{m-n} \cdot \left(1 + \frac{n}{k+\alpha^2}\right) \cdot \|\mathbf{B} - \mathbf{B}_n\|_{\mathrm{F}}^2.$$

Using $\sigma_1(\mathbf{B}) = \sigma_2(\mathbf{B}) = \cdots = \sigma_n(\mathbf{B}) = n + \alpha^2$; $\sigma_{n+1}(\mathbf{B}) = \sigma_{n+2}(\mathbf{B}) = \cdots = \sigma_m = \alpha^2$; $\|\mathbf{B} - \mathbf{B}_n\|_2^2 = \alpha^2$; and $\|\mathbf{B} - \mathbf{B}_n\|_{\mathrm{F}}^2 = (m-n)\alpha^2$, we obtain,

$$\|\mathbf{B} - \mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^{\dagger}\mathbf{B}\|_{\mathrm{F}}^2 = (m-k) \cdot \left(1 + \frac{n}{k+\alpha^2}\right) \cdot \|\mathbf{B} - \mathbf{B}_n\|_2^2.$$

Now, Lemma 7 of [10] implies that, for any matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$, rank parameter $n < \mathrm{rank}(\mathbf{W})$, and sampling parameter $n \leq k \leq m$, for any $\mathcal{S}$ of cardinality $k$, if $\mathbf{Z}_{\mathcal{S}}$ has full rank ($\mathbf{Z}$ is defined shortly),

$$\|\mathbf{W} - \mathbf{W}_{\mathcal{S}}\mathbf{W}_{\mathcal{S}}^{\dagger}\mathbf{W}\|_{\mathrm{F}}^2 \leq \|\mathbf{W} - \mathbf{W}_n\|_{\mathrm{F}}^2 + \|\,(\mathbf{W} - \mathbf{W}_n)_{\mathcal{S}}\,\mathbf{Z}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2.$$

Here, $\mathbf{W}_n$ is the best rank $n$ approximation to $\mathbf{W}$. $\mathbf{Z}$ is defined as follows. Let $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be the SVD of $\mathbf{W}$. $\mathbf{Z}$ is the first $n$ rows of $\mathbf{V}^{\mathrm{T}}$. Applying spectral submultiplicativity to this relation we obtain,

$$\|\mathbf{W} - \mathbf{W}_{\mathcal{S}}\mathbf{W}_{\mathcal{S}}^{\dagger}\mathbf{W}\|_{\mathrm{F}}^2 \leq \|\mathbf{W} - \mathbf{W}_n\|_{\mathrm{F}}^2 + \|\mathbf{Z}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \cdot \|\mathbf{W} - \mathbf{W}_n\|_2^2.$$

We now apply Lemma 7 from [10] on $\mathbf{B}$ and $\mathbf{X}$ and combine it with the bound from Theorem 19 of [10],

$$\|\mathbf{X}_{\mathcal{S}}^{\dagger}\|_{\mathrm{F}}^2 \geq \frac{\|\mathbf{W} - \mathbf{W}_{\mathcal{S}}\mathbf{W}_{\mathcal{S}}^{\dagger}\mathbf{W}\|_{\mathrm{F}}^2}{\|\mathbf{W} - \mathbf{W}_n\|_2^2} - \frac{\|\mathbf{W} - \mathbf{W}_n\|_{\mathrm{F}}^2}{\|\mathbf{W} - \mathbf{W}_n\|_2^2} = (m-k)\cdot\left(1 + \frac{n}{k+\alpha^2}\right) - (m-n) = \left(\frac{m-k}{k+\alpha^2} + 1 - \frac{k}{n}\right)\cdot\|\mathbf{X}^{\dagger}\|_{\mathrm{F}}^2.$$

■

As $\alpha \to 0$ and $k = O(n)$, this bound is $m/k - O(1)$. If $k = (1 + \Omega(1))n$ the Frobenius norm bounds in Theorems 3.1 and 3.5 asymptotically match this lower bound. However, if $k = (1 + o(1))n$ there is a gap between the lower bound and the best upper bound. There is also a gap when $k = \omega(n)$. We believe that the gap for $k = \omega(n)$ is the result of looseness in the lower bound, but we were unable to prove a tighter bound than Theorem 4.2.

# 5 Low-stretch Spanning Trees and Subset Selection

Let $G = (V, E, w)$ be a weighted undirected connected graph. Unless otherwise stated, in this section, we denote the number of vertices of $G$ by $n$, and the number of edges by $m$. Let $T = (V, F, w)$ be a spanning tree of $G$, where $F$ is a subset of $E$ having exactly $n-1$ edges. We use the same weight function $w$ because the edges in $T$ have the same weights with the corresponding edges in $G$. Since $T$ is a tree, every pair of vertices in $T$ is connected by a unique path in $T$. For any edge $e \in E$, let us denote by $p_T(e)$ the set of edges on the unique path in $T$ between the incident vertices of $e$. The stretch of $e$ with respect to $T$ is $\mathrm{St}_{\mathbf{T}}(e) = \sum_{e' \in p_T(e)} \frac{w(e)}{w(e')}$. The stretch of the graph $G$ with respect to $T$ is [2]

$$\mathrm{St}_T(G) = \sum_{e \in E} \mathrm{St}_T(e) .$$

The problem of finding a low-stretch spanning tree is the problem of finding a spanning tree $T$ of $G$ such that $\mathrm{St}_T(G)$ is minimized, among all possible spanning trees of $G$. Let $\mathrm{St}(n) = \max_{G \in G_n} \min_T \mathrm{St}_T(G)$, where $G_n$ is the family of graphs with $n$ vertices. The following bounds are known: $\mathrm{St}(n) = \Omega(m \log n)$ [2]; $\mathrm{St}(n) = O(m \log n \cdot \log\log n \cdot (\log\log\log n)^3)$ [1]. In this section we show that finding a low-stretch spanning tree is in fact an instance of the Frobenius norm version of Problem 1.1.

Finding a low stretch spanning tree has quite a few uses. One important application is the solution of symmetric diagonally dominant (SDD) linear systems of equations. Boman and Hendrickson [6] were the first to suggest the use of low-stretch spanning trees to build preconditioners for SDD matrices. Spielman and Teng [51] later showed how to use low stetch spanning trees to solve SDD systems using a nearly linear amount of operations. The currently most efficient algorithm for solving SDD systems [41] uses a low stretch spanning tree as well. One of the many obstacles in generalizing these algorithms for wider classes of matrices (e.g., finite-element matrices) is the lack of an equivalent concept, like the stretch, for such matrices. By studying the purely linear-algebraic nature of the low-stretch spanning tree problem (i.e. the Frobenius norm version of Problem 1.1), our hope is to glean new insights on how to generalize the concept of low-stretch trees, or to substitute it with something else.

Other applications of low-stretch spanning trees include: Alon-Karp-Peleg-West game, MCT approximation and message-passing model. See [26] for details.

Next, we show that finding a low-stretch spanning tree is an instance of subset selection. We first relate graphs to matrices.

**Definition 5.1** (Edge-vertex incidence matrix/Laplacian matrix)**.** *Let $G = (V, E, w)$ be a weighted undirected graph. Assume, without loss of generality, that $V = \{1, 2, \ldots, n\}$; $E = \{(u_1, v_1), (u_2, v_2), \ldots, (u_m, v_m)\}$. The* edge-vertex incidence matrix *of $G$ is $\mathbf{\Pi}_G \in \mathbb{R}^{n \times m}$, where column $i$ of $\mathbf{\Pi}_G$ is $\sqrt{w(u_i, v_i)}(\mathbf{e}_{u_i} - \mathbf{e}_{v_i})$. Here $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are the identity (standard basis) vectors. The* Laplacian matrix *of $G$ is $\mathbf{L}_G = \mathbf{\Pi}_G \mathbf{\Pi}_G^{\mathrm{T}}$.*

Every column in $\mathbf{\Pi}_G$ represents an edge in $G$. A spanning tree $T$ is a group of edges that span $G$ and form a graph. The set of edges in $T$ correspond to a set of columns in $\mathbf{\Pi}_G$, which we denote by $\mathcal{S}(T)$. Notice that if the indices are kept consistently, then, $\mathbf{\Pi}_T = (\mathbf{\Pi}_G)_{\mathcal{S}(T)}$. If $\mathcal{S} \subseteq [m]$ is a subset of columns, then, there is a subgraph $H$ of $G$ that contains the edges corresponding to the columns in $\mathcal{S}$. We denote this subgraph by $H(\mathcal{S})$. We are now ready to connect low-stretch spanning trees and subset selection.

**Theorem 5.2.** *Let $G$ be a weighted undirected connected graph. Let $\mathbf{\Pi}_G = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be the SVD of $\mathbf{\Pi}_G$ with $\mathbf{U} \in \mathbb{R}^{n\times(n-1)}$, $\mathbf{\Sigma} \in \mathbb{R}^{(n-1)\times(n-1)}$ and $\mathbf{V} \in \mathbb{R}^{m\times(n-1)}$ ($G$ is connected; so, $\mathrm{rank}(\mathbf{\Pi}_G) = n-1$). For notational convenience, let $\mathbf{Y} = \mathbf{V}^{\mathrm{T}}$.*

1. *If $T$ is a spanning tree of $G$, then, $\mathrm{St}_T(G) = \|\mathbf{Y}_{\mathcal{S}(T)}^{-1}\|_{\mathrm{F}}^2$.*
2. *If $\mathcal{S} \subseteq [m]$ has cardinality $n-1$ and $\mathbf{Y}_{\mathcal{S}}$ has full rank, then, $H(\mathcal{S})$ is a spanning tree of $G$.*

*Proof.* To prove the first part of Theorem 5.2, we need a result of Spielman and Woo [50], who recently connected the stretch of $G$ with respect to $T$ to the matrix $\mathbf{L}_G\mathbf{L}_T^{\dagger}$. More precisely, Theorem 2.1 in [50] shows that if $T$ is a spanning tree then $\mathrm{St}_T(G) = \mathrm{Tr}\left(\mathbf{L}_G\mathbf{L}_T^{\dagger}\right)$. Here, $G$ is a weighted undirected connected graph and $T$ is a spanning tree of $G$.

Let us denote $\mathcal{S} = \mathcal{S}(T)$. Since $T$ is a tree we have $\mathrm{rank}(\mathbf{\Pi}_T) = n-1$ (it is well known that the edge incidence matrix of a connected graph has rank $|V|-1$). Since $\mathbf{\Pi}_T = (\mathbf{\Pi}_G)_{\mathcal{S}} = \mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}}$, $\mathbf{Y}_{\mathcal{S}}$ must be full rank. Now,

$$\begin{aligned}
\mathrm{St}_T(G) \overset{(a)}{=} \mathrm{Tr}\left(\mathbf{L}_G\mathbf{L}_T^{\dagger}\right) \overset{(b)}{=} \mathrm{Tr}\left(\mathbf{\Pi}_G\mathbf{\Pi}_G^{\mathrm{T}}\left((\mathbf{\Pi}_G)_{\mathcal{S}}(\mathbf{\Pi}_G)_{\mathcal{S}}^{\mathrm{T}}\right)^{\dagger}\right) &\overset{(c)}{=} \mathrm{Tr}\left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^{\mathrm{T}}(\mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}}\mathbf{Y}_{\mathcal{S}}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}})^{\dagger}\right) \\
&\overset{(d)}{=} \mathrm{Tr}\left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}^{-1}\left(\mathbf{Y}_{\mathcal{S}}\mathbf{Y}_{\mathcal{S}}^{\mathrm{T}}\right)^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}\right) \\
&\overset{(e)}{=} \mathrm{Tr}\left(\mathbf{\Sigma}\left(\mathbf{Y}_{\mathcal{S}}\mathbf{Y}_{\mathcal{S}}^{\mathrm{T}}\right)^{-1}\mathbf{\Sigma}^{-1}\right) \\
&\overset{(f)}{=} \mathrm{Tr}\left(\left(\mathbf{Y}_{\mathcal{S}}\mathbf{Y}_{\mathcal{S}}^{\mathrm{T}}\right)^{-1}\right) \\
&= \|\mathbf{Y}_{\mathcal{S}(T)}^{-1}\|_{\mathrm{F}}^2 .
\end{aligned}$$

$(a)$ follows by the Spielman-Woo result. $(b)$ follows by replacing the Laplacian matrices with the product of their edge-incidence matrices. $(c)$ follows by introducing the SVD of $\mathbf{\Pi}_G$ and the equality $(\mathbf{\Pi}_G)_{\mathcal{S}} = \mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}}$. $(d)$ follows since all three matrices involved ($\mathbf{U}$, $\mathbf{Y}_{\mathcal{S}}$, and $\mathbf{\Sigma}$) are full rank. $(e)$ follows since $\mathbf{U}$ has orthonormal columns. $(f)$ follows since $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$.

We now prove the second part of the theorem. For $H(\mathcal{S})$ to be a spanning tree, it has to be a connected graph with $n-1$ edges. The last condition is met since $\mathcal{S}$ has cardinality $n-1$, and the number of edges in $H(\mathcal{S})$ is equal to the cardinality of $\mathcal{S}$. As for connectivity, notice that $\mathbf{\Pi}_{H(\mathcal{S})} = (\mathbf{\Pi}_G)_{\mathcal{S}} = \mathbf{U}\mathbf{\Sigma}\mathbf{Y}_{\mathcal{S}}$. Now, since $\mathbf{Y}_{\mathcal{S}}$ has full rank, we have $\mathrm{rank}(\mathbf{\Pi}_{H(\mathcal{S})}) = |V|-1$. This directly implies that $H(\mathcal{S})$ is connected. ■

The algorithms that we presented in Theorems 3.1 and 3.11 in Section 3 can be used to find a low-stretch spanning tree (run these algorithms on the matrix $\mathbf{Y}$ of the above theorem), but they are not competitive both in terms of operation count and in terms of approximation bounds. Both these algorithms can guarantee $\mathrm{St}_T(G) \leq (n-1)(m-n+2)$. The operation count is $(m^2n)$ and $O(mn^3)$, respectively. This upper bound also holds for the easily computable maximum weight spanning tree. Koutis et al. describe in [41] an algorithm which gives the available state-of-the-art upper bound $\mathrm{St}_T(G) \leq O(m\log n \cdot \log\log n \cdot (\log\log\log n)^3)$, and has operation count of $O(m\log(n) + n\log(n)\log\log(n))$. The main reason for this gap is that our algorithms are designed for general matrices, while [1, 41] describe a graph algorithm, which better exploits the unique structure of the problem. Nevertheless, when reinterpreting our algorithms as algorithms for constructing low-stretch spanning trees yields interesting connections that sheds light on both problems, as we discuss below.

### 5.1 Low-stretch spanning trees via the greedy removal algorithm

Theorem 5.2 along with Theorem 3.1 suggest a greedy removal algorithm for constructing a spanning tree with low stretch: start with a full set of edges $H = E$; then, at each iteration, find the edge $e$ such that

$\|\mathbf{Y}^{\dagger}_{\mathcal{S}(H-\{e\})}\|_{\mathrm{F}}^2$ is minimized, and set $H \longleftarrow H-\{e\}$. Finish once $H$ has $n-1$ edges. That is, we apply the algorithm of Theorem 3.1 on $\mathbf{Y}$ (see Theorem 5.2 for the definition of $\mathbf{Y}$). We note that this algorithm is *different* from the natural greedy removal algorithm, which would remove edges to keep the stretch of the subgraph minimal in each step. It is possible to define the stretch $\mathrm{St}_H(G)$ of a subgraph $H$; we refer the reader to chapter 18 of [46] for the definition. It is also possible to show that for a spanning subgraph $H$, $\|\mathbf{Y}^{\dagger}_{\mathcal{S}(H)}\|_{\mathrm{F}}^2 \leq \mathrm{St}_H(G)$ (we omit the proof), but an equality does not hold. In fact, our algorithmic results imply that it is possible to find a subgraph with $O(n)$ edges such that $\|\mathbf{Y}^{\dagger}_{\mathcal{S}(H)}\|_{\mathrm{F}}^2 = O(m)$, but there exists a graph for which every subgraph $H$ of $O(n)$ edges we have $\mathrm{St}_H(G) = \Omega(m\log(n))$ (Corollary 18.1.5 in [46]).

We conducted some simple experiments with our greedy removal algorithm. In the first experiment, we used greedy removal to generate a spanning tree $T_n$ of the complete graph $K_n$ on $n$ with equal weights vertices, for $n = 10, 11, \ldots, 50$. We then computed the stretch of $T_n$. We found that $\mathrm{St}_{T_n}(K_n) \approx 0.6m\log^2 n$. We then repeated this experiment with random weights on the edges of $K_n$. We found that in almost all runs, $\mathrm{St}_{T_n}(K_n) \approx 0.3m\log^2 n$. These values are much better than our theoretical bounds, and are closer to what it is possible to find using state-of-the-art algorithms for low-stretch trees. These experiments, although far from exhaustive, suggest that our theoretical *worst-case* upper bounds for greedy removal are rather pessimistic for the matrices relevant to finding a low-stretch spanning tree.

### 5.2 Maximum weight spanning trees and maximum volume subsets

The volume corresponding to a set $\mathcal{S}$ has a very natural interpretation when $\mathcal{S}$ is a subset of columns in $\mathbf{\Pi}_G$, and it corresponds to a tree. Let $\mathbf{\Pi}_G = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be the SVD of $\mathbf{\Pi}_G$ ($\mathbf{U} \in \mathbb{R}^{n\times(n-1)}$, $\mathbf{\Sigma} \in \mathbb{R}^{(n-1)\times(n-1)}$, and $\mathbf{V} \in \mathbb{R}^{m\times(n-1)}$; $G$ is connected so $\mathrm{rank}(\mathbf{\Pi}_G) = n-1$). For notational convenience, let $\mathbf{Y} = \mathbf{V}^{\mathrm{T}}$. Define $\bar{\mathbf{U}}$ to be the first $n-1$ rows of $\mathbf{U}$; $\bar{\mathbf{U}}$ is a square matrix). Define $\bar{\mathbf{\Pi}}_G$ to be the first $n-1$ rows of $\mathbf{\Pi}_G$. Notice that $\bar{\mathbf{\Pi}}_G = \bar{\mathbf{U}}\mathbf{\Sigma}\mathbf{Y}$. This implies that $S_k(\mathbf{Y}) = S_k(\bar{\mathbf{\Pi}}_G)$ and also that the set $\mathcal{S}$ that maximizes $\det(\mathbf{Y}_{\mathcal{S}})^2$ also maximizes $\det((\bar{\mathbf{\Pi}}_G)_{\mathcal{S}})^2$.

Let $T$ be a spanning tree of $G$. Determinants of the form $\det((\bar{\mathbf{\Pi}}_G)_{\mathcal{S}(T)}(\bar{\mathbf{\Pi}}_G)^{\mathrm{T}}_{\mathcal{S}(T)})$ have a very natural interpretation. The matrix $(\bar{\mathbf{\Pi}}_G)_{\mathcal{S}(T)}(\bar{\mathbf{\Pi}}_G)^{\mathrm{T}}_{\mathcal{S}(T)}$ is a Laplacian of a graph for which a column and row of some vertex have been removed. It is well known that the determinant of such matrices, when the graph is a tree, is equal to the product of the weights of the tree edges. That is,

$$\det((\bar{\mathbf{\Pi}}_G)_{\mathcal{S}(T)}(\bar{\mathbf{\Pi}}_G)^{\mathrm{T}}_{\mathcal{S}(T)}) = \prod_{e\in T} w(e)\,.$$

The subset of columns $\mathcal{S}$ that maximizes the volume also maximizes $\prod_{e\in H(\mathcal{S})} w(e)$. This trivially implies that the corresponding tree is a maximum weight spanning tree. So, for edge-incidence matrices one can use an efficient maximum weight spanning tree algorithm to find the maximum volume subset of columns efficiently. The bound we obtain is $\mathrm{St}_T(G) < (n-1)(m-n+2)$. We are unaware of any other analysis of the stretch of a maximum weight spanning tree, but this bound can be easily proven using much simpler arguments.

### 5.3 Low-stretch spanning trees via volume sampling

Recall Problem 1.1, and let the input matrix be the matrix $\mathbf{Y}$ from Theorem 5.2. Using volume sampling (Lemma 3.9) to sample a subset of columns from this $\mathbf{Y}$ corresponds to sampling a random spanning tree, where a tree $T$ is sampled with relative probability $\prod_{e\in T} w(e)$. We denote this probability distribution on spanning trees of $G$ by $\Gamma(G)$. Lemma 3.9 provides a bound on the stretch of a random spanning tree sampled from $\Gamma(G)$. Notice that it is a strict upper bound. The reason is that not every subgraph $H$ is a tree. We conjecture that this bound is pessimistic, and leave for future work the refinement of the bound.

**Corollary 5.3.** *Let $G$ be a weighted undirected connected graph, and let $\mathcal{T}$ be a random spanning tree, where tree $T$ is sampled with relative probability $\prod_{e\in T} w(e)$. Then,*

$$\mathbb{E}\left[\mathrm{St}_{\mathcal{T}}(G)\right] < (n-1)(m-n+2)\,.$$

One can use VolumeSample from [34] to generate such a spanning tree in $O(n^3m)$ operations. However, the problem of generating a sample from $\Gamma(G)$ is a well studied problem, and there exists algorithms that can generate a random spanning tree faster than $O(n^3m)$. See [56] for a short review.

### 5.4 Towards better bounds for low-stretch spanning trees

State-of-the-art algorithms for finding low stretch spanning trees attain theoretical worst-case bounds that are better than the ones we obtain for a general matrix. We now provide a preliminary explanation for this gap.

Consider a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ with orthonormal rows such that for every subset $\mathcal{S} \subseteq [m]$ of cardinality $n$, $\det(\mathbf{X}_{\mathcal{S}})^2 = 0$ or $\det(\mathbf{X}_{\mathcal{S}})^2 = C$, for some constant $C$. The second inequality of Lemma 3.8 is

$$\|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2 = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\det\left(\mathbf{X}_{\mathcal{S}}(i\to\mathbf{x}_j)\right)^2}{\det\left(\mathbf{X}_{\mathcal{S}}\right)^2}.$$

(here it is an equality because $\mathbf{X}$ is orthonormal; also, $\|\mathbf{X}^{\dagger}\|_2^2 = 1$). Since all determinants are 0 or $C$, we find that

$$\|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2 = \#\{\mathcal{T} \,:\, \mathrm{rank}(\mathbf{X}_{\mathcal{T}}) = n, \mathcal{T} = (\mathcal{S}-\{i\})\cup\{j\} \text{ for } i\in\mathcal{S}, j\in[m]\}\,.$$

That is, for a subset $\mathcal{S}$ such that the columns of $\mathbf{X}$ in $\mathcal{S}$ form a basis for the column space of $\mathbf{X}$, $\|\mathbf{X}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2$ is equal to the number of bases that can be obtained by replacing a single column. The last quantity can only be bounded universally by $n(m-n+1)$, and that quantity is obtained for all $\mathcal{S}$s if $\mathbf{X}_{\mathcal{T}}$ has full rank for *every* subset $\mathcal{T}$. However, if there exist at least one subset $\mathcal{T}$ for which $\mathbf{X}_{\mathcal{T}}$ is singular then there is a subset $\mathcal{S}$ for which the $n(m-n+1)$ bound is strict. If many such sets exist, then, the bound is probably very loose.

Now, let us consider the incidence matrix of a complete graph with equal weights. If for a subset $\mathcal{S}$ the subgraph $H(\mathcal{S})$ is not a tree, then, $\det(\mathbf{Y}_{\mathcal{S}}) = 0$ ($\mathbf{Y}$ is defined in Theorem 5.2) . Every tree has exactly the same weight, so for all $\mathcal{S}$'s that correspond to trees we have the same $\det(\mathbf{Y}_{\mathcal{S}})^2$. We see that $\mathbf{Y}$ falls into the case discussed above. We conclude that the reason that $\mathbf{Y}$ has a subset of column $\mathcal{S}$ for which $\|\mathbf{Y}_{\mathcal{S}}^{-1}\|_{\mathrm{F}}^2$ is small is the fact that for some subsets $\mathcal{T}$ the matrix $\mathbf{Y}_{\mathcal{T}}$ does not have full rank. In fact, for the complete graph, most cardinality $n-1$ subsets of edges will not result in a tree or a full rank $\mathbf{Y}_{\mathcal{S}}$. If we could enumerate these subsets exactly, this should give a better upper bound for this special matrix.

## 6 Other Applications

### 6.1 Column-Based Low-Rank Matrix Reconstruction

Suppose we want to build a low rank approximation of $\mathbf{A} \in \mathbb{R}^{d\times m}$. For a rank parameter $r < \mathrm{rank}(\mathbf{A})$, let $\mathbf{A}_r \in \mathbb{R}^{d\times m}$ denote the best rank $r$ approximation to $\mathbf{A}$. That is, $\mathbf{A}_r$ minimizes $\|\mathbf{A}-\mathbf{B}\|_2$, over $\mathbf{B}$, where $\mathbf{B}$ ranges on all rank $r$ $d\times m$ matrices. It is well known that $\mathbf{A}_r$ can be computed via the SVD of $\mathbf{A}$. However, SVD uses all the columns of $\mathbf{A}$ to compute $\mathbf{A}_r$. In some applications it is desirable to use only a small set of columns to build the low rank approximation (see [10] and references there in for such applications). Let $\mathcal{S} \subseteq [m]$ and $\mathbf{A}_{\mathcal{S}}$ contains a subset of columns of $\mathbf{A}$ indicated in $\mathcal{S}$. Define $\mathbf{\Pi}_{\mathcal{S},r}(\mathbf{A}) \in \mathbb{R}^{d\times m}$ to be the best rank $r$ approximation of $\mathbf{A}$ within the columns space of $\mathbf{A}_{\mathcal{S}}$, with respect to the spectral norm

(if $\mathcal{S} = [m]$, then $\mathbf{\Pi}_{\mathcal{S},r}(\mathbf{A}) = \mathbf{A}_r$). The so-called column-based low-rank matrix reconstruction problem is: given $\mathbf{A}$, $r < \mathrm{rank}(\mathbf{A})$, and an sampling parameter $k \geq r$, find a subset $\mathcal{S}$ of cardinality at most $k$ such that $\|\mathbf{A} - \mathbf{\Pi}_{\mathcal{S},r}(\mathbf{A})\|_2$ is minimized among all the possible choices for the subset $\mathcal{S}$.

It is natural to evaluate $\mathbf{\Pi}_{\mathcal{S},r}(\mathbf{A})$ in terms of $\mathbf{A}_r$. That is, provide approximation bounds of the form $\|\mathbf{A} - \mathbf{\Pi}_{\mathcal{S},r}(\mathbf{A})\|_2 \leq \alpha \cdot \|\mathbf{A} - \mathbf{A}_r\|_2$. Currently, the best deterministic such algorithms are available in [10]. These algorithms achieve asymptotically optimal upper bounds, but there is still room for improvement in terms of lowering the the operation count.

The algorithm of Corollary 3.3 can be used to obtain a new deterministic algorithm for the column-based low-rank matrix reconstruction problem. First, construct an SVD decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$. Let $\mathbf{X} \in \mathbb{R}^{r \times m}$ be the first $r$ rows of $\mathbf{V}^{\mathrm{T}}$. We now use the algorithm of Corollary 3.3 on $\mathbf{X}$ to generate a subset $\mathcal{S} \subseteq [m]$ of size $k$, which is the result of the algorithm. The following bound holds,

$$\|\mathbf{A} - \mathbf{\Pi}_{\mathcal{S},r}(\mathbf{A})\|_2 \leq \sqrt{2 + \frac{r(m-k)}{k-r+1}} \cdot \|\mathbf{A} - \mathbf{A}_r\|_2 \,.$$

We omit the proof, which follows by combining Lemma 7 from [9] and Corollary 3.3. The algorithm is deterministic and the operation count is $T_{SVD} + O(mr(m-k))$, where $T_{SVD}$ is the number of operations needed to compute the top $r$ right singular vectors of $\mathbf{A}$. Our approximation bound is slightly worse than the bounds in [10] but bound on the number of operations can sometimes be better, depending on the size of the input matrix. We refer the interested reader to [10] to conduct her own comparison.

### 6.2 Sparse Solutions to Least-squares Regression Problems

Fix inputs $\mathbf{A} \in \mathbb{R}^{d \times m}$ and $\mathbf{b} \in \mathbb{R}^d$; consider the following least-squares problem, $\min_{\mathbf{x} \in \mathbb{R}^m} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$. Since there is no assumption on $d$ and $m$, or that $\mathbf{A}$ is full rank, the minimizer of $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ might not be unique; there might be a full subspace of minimizers. Even if there is a unique minimizer, it might have a huge norm, while there exists an almost-minimizer with small norm. It depends on the application what exactly is needed, but often some kind of *regularization* is used to address the issues just mentioned. One popular regularization technique is *truncated* SVD [35]: for $r < \mathrm{rank}(\mathbf{A})$, let $\mathbf{A}_r \in \mathbb{R}^{d \times m}$ of rank $r$ denote the rank-$r$ SVD of $\mathbf{A}$; then, the truncated SVD regularized solution is given by $\mathbf{x}_{svd(r)} = \mathbf{A}_r^{\dagger}\mathbf{b} \in \mathbb{R}^m$.

However, sometimes a different regularization is sought: requiring the solution vector to be sparse. That is, we are interested in constructing a vector $\mathbf{x}_k \in \mathbb{R}^m$ that has at most $k$ non-zeros, for some $k$. Since truncated SVD is arguably the most natural regularizer, it makes sense to compare $\mathbf{x}_k$ to $\mathbf{x}_{svd(r)}$ for some $r \leq k$. More specifically, we are interested in bounds of the form,

$$\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|_2 \leq \|\mathbf{A}\mathbf{x}_{svd(r)} - \mathbf{b}\|_2 + \alpha \,.$$

The idea of obtaining sparse solutions with approximation bounds of the above type can be traced to [15]. Currently, the best *deterministic* method is in [8] ($k > r$) with $\alpha = \left(1 + \sqrt{\frac{r}{k}}\right) \|\mathbf{b}\|_2 \|\mathbf{A} - \mathbf{A}_r\|_{\mathrm{F}} / \sigma_r(\mathbf{A})$.

The algorithm of Corollary 3.3 can be used to design a new deterministic algorithm. First, construct an SVD decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$. Let $\mathbf{X} \in \mathbb{R}^{r \times m}$ be the first $r$ rows of $\mathbf{V}^{\mathrm{T}}$. We now use the algorithm of Corollary 3.3 on $\mathbf{X}$ to generate a subset $\mathcal{S} \subseteq [m]$ of size $k$. We now compute $\hat{\mathbf{x}}_k = \mathbf{A}_{\mathcal{S}}^{\dagger}\mathbf{b} \in \mathbb{R}^k$. We form $\mathbf{x}_k$ by spreading the indices of $\hat{\mathbf{x}}_k$ to their corresponding place in $\mathbf{x}_k$ and putting 0 elsewhere. The following bound holds,

$$\|\mathbf{A}\mathbf{x}_k - \mathbf{b}\|_2 \leq \|\mathbf{A}\mathbf{x}_{svd(r)} - \mathbf{b}\|_2 + \left(1 + \sqrt{\frac{r(m-k)}{k-r+1}}\right) \|\mathbf{b}\|_2 \frac{\sigma_{r+1}(\mathbf{A})}{\sigma_r(\mathbf{A})} \,.$$

We omit the proof since it follows immediately by combining Lemma 3 from [8] with Corollary 3.3. The algorithm is deterministic and the operation count is $O(dm \min\{d, m\} + mr(m-k))$.

Our bound essentially contains the term $\sqrt{m-k} \cdot \sigma_{r+1}(\mathbf{A})$ in place of the term $\|\mathbf{A} - \mathbf{A}_r\|_\mathrm{F}$ in the bound of [8]. It is always the case that $\|\mathbf{A} - \mathbf{A}_r\|_\mathrm{F} \leq \sqrt{m-r} \cdot \sigma_{r+1}(\mathbf{A})$, but since $k \geq r$, our bound might be better in some cases (e.g. when $k \to m$).

### 6.3 Feature Selection in $k$-Means Clustering

The deterministic algorithm of Corollary 3.3 can also be used for deterministic feature selection in $k$-means clustering. We refer the reader to [11] for an introduction to this problem. Theorem 4 of [11] gives such a polynomial deterministic unsupervised feature selection algorithm, which selects features from the data and then *rescales* them. Using Corollary 3.3, one can design a deterministic unsupervised feature selection algorithm *without* rescaling. We omit the details, since the algorithm is similar to the one described for sparse least squares, and the analysis is a combination of Lemma 10 from [11] with Corollary 3.3. The approximation bound that is obtained is $O(r(m-k)/(k-r+1))$; here, $m$ is the number of features in the input points, $r$ is the number of clusters, and $k$ is the number of features after feature selection.

## 7 Open Problems and Future Directions

Several interesting questions remain unanswered and we leave them for future investigation. First, is the Frobenius-norm version of Problem 1.1 NP-hard? Second, is it possible to close the existing gaps between lower and upper bounds for Problem 1.1? Third, is it possible to extend the Strong Rank Revealing QR method of [32] to sample arbitrary $k \geq n$ columns? Fourth, is it possible to extend the polynomial implementations of volume sampling in [22, 34] to sample arbitrary number of columns from short-fat matrices? Finally, is it possible to derandomize the algorithm of Theorem 3.11?

## Acknowledgements

We would like to thank Ioannis Koutis for bringing [44, 40, 41] to our attention; Petros Drineas, Frank De Hoog, Sivan Toledo, and Mark Tygert for many useful discussions and suggestions on a preliminary draft of this work; and Anastasios Zouzias for pointing out the connection between the restricted invertibility line of research [7, 54, 49] and ours.

## References

[1] I. Abraham, Y. Bartal, and O. Neiman. Nearly tight low stretch spanning trees. In *Proceedings of the 49th IEEE symposium on Foundations of Computer Science (FOCS)*, 2008.

[2] N. Alon, R. M. Karp, D. Peleg, and D. West. A graph-theoretic game and its application to the k-server problem. *SIAM J. Comput.*, 24:78–100, February 1995.

[3] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

[4] J. Batson, D. Spielman, and N. Srivastava. Twice-ramanujan sparsifiers. In *Proceedings of the 41st annual ACM symposium on Theory of Computing (STOC)*, 2009.

[5] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas with Applications to Linear Systems Theory.* Princeton University Press, 2005.

[6] E. G. Boman and B. Hendrickson. On spanning tree preconditioners. Unpublished manuscript, Sandia National Laboratories, 2001.

[7] J. Bourgain and L. Tzafriri. Invertibility of large submatrices with applications to the geometry of banach spaces and harmonic analysis. *Israel Journal of Mathematics*, 57:137–224, 1987.

[8] C. Boutsidis. On truncated-svd-like sparse solutions to least-squares of arbitrary dimensions. *Manuscript*. `http://www.cs.rpi.edu/~boutsc/SparseChristos.pdf`.

[9] C. Boutsidis. Topics in matrix sampling algorithms. *PhD Thesis, Rensselaer Polytechnic Institute*, May, 2011.

[10] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near optimal column based matrix reconstruction. In *Proceedings of the 52hd Annual IEEE Symposium on Foundations of Computer Science (FOCS). Full version available on line: `http://arxiv.org/pdf/1103.0995v2.pdf`; also invited to SICOMP special issue*, 2011.

[11] C. Boutsidis and M. Magdon-Ismail. Deterministic feature selection for $k$-means clustering. *Arxiv Preprint: 1109:5664*, Sept, 2011.

[12] C. Boutsidis, M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the 19th Annual ACM -SIAM Symposium on Discrete Algorithms (SODA)*, pages 968–977, 2009.

[13] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas. Stochastic dimensionality reduction for $k$-means clustering. *Arxiv Preprint: 1110:2897*, Oct, 2011.

[14] M. Broadbent, M. Brown, K. Penner, I. Ipsen, and R. Rehman. Subset selection algorithms: Randomized vs. deterministic. *SIAM Undergraduate Research Online*, 3, 2010.

[15] T. Chan and P. Hansen. Some applications of the rank revealing QR factorization. *SIAM Journal on Scientific and Statistical Computing*, 13:727, 1992.

[16] T. F. Chan and P. C. Hansen. Some applications of the rank revealing QR factorization. *SIAM Journal on Scientific and Statistical Computing*, 13:727–741, 1992.

[17] T. F. Chan and P. C. Hansen. Low-rank revealing QR factorizations. *Numerical Linear Algebra with Applications*, 1:33–44, 1994.

[18] A. Civril and M. Magdon-Ismail. Exponential inapproximability of selecting a maximum volume sub-matrix. *Algorithmica*. to appear.

[19] A. Civril and M. Magdon-Ismail. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410:4801–4011, 2009.

[20] F. de Hoog and R. Mattheij. Subset selection for matrices. *Linear Algebra Appl.*, 422:349–359, 2007.

[21] F. de Hoog and R. Mattheij. A note on subset selection for matrices. *Linear Algebra Appl.*, 434:1845–1850, 2011.

[22] A. Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 42th Annual ACM Symposium on Theory of Computing (STOC)*, 2010.

[23] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1117–1126, 2006.

[24] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. In *International Conference on Machine Learning (ICML)*, 2012.

[25] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlos. Faster least squares approximation. *Numer. Math.*, 117(2):219–249, Feb. 2011.

[26] M. Elkin, Y. Emerk, D. Spielman, and S. Teng. Lower-stretch spanning trees. In *Proceedings of the 37th ACM symposium on Theory of computing (STOC)*, 2005.

[27] L. Foster and R. Kommu. Algorithm 853: An efficient algorithm for solving rank-deficient least squares problems. *ACM Transactions on Mathematical Software (TOMS)*, 32(1):157–165, 2006.

[28] A. Gittens. The spectral norm error of the naive nystrom extension. *Arxiv preprint arXiv:1110.5305*, November, 2011.

[29] G. Golub and C. V. Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.

[30] S. Goreinov, E. Tyrtyshnikov, and N. Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra and Its Applications*, 261:1–21, 1997.

[31] M. Gu and S. Eisenstat. Downdating the singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 16(3):793–810, 1995.

[32] M. Gu and S. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17:848–869, 1996.

[33] M. Gu and L. Miranian. Strong rank revealing Cholesky factorization. *Electronic Transactions on Numerical Analysis*, 17:76–92, 2004.

[34] V. Guruswami and A. Kemal Sinop. Optimal column-based low-rank matrix reconstruction. *Technical Reprot, http://arxiv.org/abs/1104.1732*, 2011.

[35] P. Hansen. The truncated svd as a method for regularization. *BIT Numerical Mathematics*, 27(4):534–553, 1987.

[36] Y. P. Hong and C. T. Pan. Rank-revealing QR factorizations and the singular value decomposition. *Mathematics of Computation*, 58:213–232, 1992.

[37] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

[38] I. Ipsen, C. T. Kelley, and S. R. Pope. Rank-deficient nonlinear least squares problems and subset selection. *SIAM Journal on Numerical Analysis*, 49(3):1244–1266, 2011.

[39] S. Joshi and S. Boyd. Sensor Selection via Convex Optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.

[40] I. Koutis. Parameterized complexity and improved inapproximability for computing the largest j-simplex in a v-polytope. *Inf. Process. Lett.*, 100:8–13, October 2006.

[41] I. Koutis, G. L. Miller, and R. Peng. Solving SDD linear systems in time $\tilde{O}(m \log n \log(1/\epsilon))$. In *Proceedings of the 52st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2011.

[42] M. Magdon-Ismail. Row sampling for matrix algorithms via a non-commutative bernstein bound. *Preprint, Available online, arXiv:1008.0587*, 2010.

[43] P. Martinsson, V. Rokhlin, and M. Tygert. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30(1):47–68, 2011.

[44] A. Packer. Np - hardness of largest contained and smallest containing simplices for v- and h-polytopes. *Discrete and Computational Geometry*, 28(3):349–377, 2002.

[45] C. T. Pan. On the existence and computation of rank-revealing LU factorizations. *Linear Algebra and its Applications*, 316:199–222, 2000.

[46] D. Peleg. *Distributed computing: a locality-sensitive approach*. Society for Industrial and Applied Mathematics, 2000.

[47] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *JACM: Journal of the ACM*, 54, 2007.

[48] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.

[49] D. Spielman and N. Srivastava. An elementary proof of the restricted invertibility theorem. *Israel Journal of Mathematics*, pages 1–9.

[50] D. Spielman and J. Woo. A note on preconditioning by low stretch spanning trees. *Arxiv preprint arXiv:0903.2816*, November, 2009.

[51] D. A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th annual ACM symposium on Theory of computing (STOC)*, 2004.

[52] N. Srivastava and D. Spielman. Graph sparsifications by effective resistances. In *Proceedings of the 40th ACM Symposium on Theory of Computing (STOC)*, 2008.

[53] J. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Adv. Adapt. Data Anal., special issue, "Sparse Representation of Data and Images*, 2011.

[54] J. A. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, pages 978–986, 2009.

[55] R. Vershynin. A note on sums of independent random matrices after ahlswede-winter. *Lecture notes*, 2010. available online `http://www-personal.umich.edu/~romanv/teaching/reading-group/ahlswede-winter.pdf`.

[56] D. B. Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the 28th annual ACM symposium on Theory of computing (STOC)*, 1996.

[57] A. Zouzias. A Matrix Hyperbolic Cosine Algorithm and Applications. *Arxiv preprint arXiv:1103.2793*, 2011.